DOCUMENT RESUME

ED 161 041 CS 204 415

AUTHOR	Rhode, Mary; Cronnell, Bruce
TITLE	Compilation of a Communication Skills Lexicon Coded with Linguistic Information. Technical Report No. 58.
INSTITUTION	Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	30 Nov 77
CONTRACT	NE-0-00-3-0064
NOTE	111p.
EDRS PRICE	MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS	*Communication Skills; *Dictionaries; Elementary Education; English Instruction; Language Research; *Lexicography; Linguistics; Phonics; *Word Lists

ABSTRACT

The production of a basic kindergarten through grade six word set, with accompanying linguistic data for use in the development of elementary communication skills instruction, is detailed in this report. Section one outlines the compilation of the 10,000 word lexicon, including a discussion of the form of the entries, the sources for the entry and general lexicons, and the empirical verification study. The second section describes the codes used with the lexicon to provide linguistic information, and the third section reports on the analysis of the linguistic information obtained through the codes. Numerous tables accompany the text. Appendixes contain a report of a new word study, pronunciation symbols and notational devices used in the report, three codes for spelling-to-sound correspondences, codes for prefixes and suffixes, bound bases, and six tables of frequencies of correspondences, affixes and epenthetic elements. (FL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED161041

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

TECHNICAL REPORT 58

November 30, 1977

COMPILATION OF A COMMUNICATION SKILLS LEXICON CODED WITH LINGUISTIC INFORMATION

Mary Rhode and Bruce Cronnell

ABSTRACT

A 10,000 word lexicon, composed primarily of uninflected forms, was compiled to provide a basic vocabulary for use in developing K-6 elementary communication skills instruction. The words were drawn from eight sources according to inclusion-exclusion criteria that were based on the nature of the sources and on the forms of lexical entries. In addition, a study of new words was conducted and 300 more recent terms were added. Also included are personal and geographic names.

Each word in the lexicon was coded to provide maximal linguistic information for use in research and design. Information coded in the corpus includes: grade levels, frequencies, numbers of syllables and kinds of stress patterns, descriptions of compounds, kinds of proper nouns, affixation processes, spelling-to-sound and sound-to-spelling correspondences, affixes, and bases. The data were processed by computer.

This report lists all codes used and provides frequencies for most codes. The results of the coding are analyzed and discussed.

This document has been distributed to a limited audience for a limited purpose. It is not published. Copies may be made only with the written permission of SWRL Educational Research and Development, 4665 Lampson Avenue, Los Alamitos, California 90720. The work upon which this document is based was performed pursuant to Contract NE-0-00-3-0064 with the National Institute of Education. SWRL reports do not necessarily reflect the opinions or policies of the sponsors of SWRL R&D.

CS204415

Contents

	Page
Abstract	i
List of Tables	v
List of Illustrations	vii
INTRODUCTION	1
Section I COMPILATION	3
The Form of Lexicon Entries	3
Sources for the Entry Lexicon	5
Sources for the General Lexicon	7
Empirical Verification Study	11
Section II CODING	13
Line 1 Codes	13
Line 2 Codes	22
Coding and Computer Processing	28
Section III ANALYSIS	31
Line 1 Codes	31
Line 2 Codes	50
CONCLUSION	61
References	63
Appendices	
A New Word Study	67
B Symbols	71
C Codes for Spelling-to-Sound Correspondences	73
D Codes for Exceptional Spelling-to-Sound Correspondences	87
E Additional Codes for Spelling-to-Sound Correspondences	89
F Codes for Prefixes	93
G Codes for Suffixes	97

Page

H Bound Bases 101

I Frequencies of Correspondences, Affixes, and Epenthetic Elements 103

Table 1 Frequencies of Consonants
Table 2 Frequencies of Secondary Vowels
Table 3 Frequencies of Primary Vowels
Table 4 Frequencies of Prefixes
Table 5 Frequencies of Suffixes
Table 6 Frequencies of Epenthetic Elements

List of Tables

Table		Page
1	Summary of Lexicon Sources	6
2	Inclusion Criteria for Geographic Names	10
3	Computer Arrangement of Codes	14
4	Grade Leveling Criteria	17
5	Codes for Affixation Processes	23
6	Percentages of Low to High Frequency Words by Grade Level	35
7	Grade Level Frequencies	36
8	Frequencies of Syllables and Stress Patterns	37
9	Frequencies of Compounds, Pseudocompounds, Contractions, and Abbreviations	39
10	Frequencies of Affix Occurrence	40
11	Frequencies of Codes about Compounds, Pseudocompounds, and Contractions	41
12	Frequencies of Miscellaneous Features	43
13	Frequencies of Names	45
14	Frequencies of Affixation Processes	46
15	Rank Order of Consonant Grapheme Units in Bases	51
16	Rank Order of Secondary Vowel Grapheme Units in Bases	52
17	Rank Order of Primary Vowel Graphemes in Bases	52
18	Rank Order of Prefixes	54
19	Rank Order of Suffixes	55
20	Rank Order of Epenthetic Elements	56
21	Rank Order of Free Bases with Frequencies Over 10	57
22	Rank Order of Bound Bases with Frequencies Over 10	59

List of Illustrations

Figure		Page
1	Sample page from computer printout.	30
2	Occurrence of low to high frequency words.	33

COMPILATION OF A COMMUNICATION SKILLS LEXICON CODED WITH LINGUISTIC INFORMATION

Mary Rhode and Bruce Cronnell

The goal of this lexical study was to provide a basic K-6 word set with accompanying linguistic data for use in the development of elementary communication skills instruction. This paper describes the compilation of the lexicon, the coding used to provide the linguistic information, and the results of the compilation and coding.

While an earlier lexicon (Berdiansky, Cronnell, & Koehler, 1969) established a useful information base relevant to K-3 reading instruction, its limitations necessitated a more comprehensive and complete compilation relevant to K-6 communication skills instruction. The current lexicon differs from the above lexicon, not only in its wider coverage (K-6 rather than K-3), but also in its more rigorous compilation methods which utilize multiple and varied sources. In addition, the revised codes for linguistic information provide more extensive data for researchers and developers.

There have been several extensive vocabulary studies, but each has limitations which restrict its value for use in inquiry concerned with elementary communication skills instruction. One of the most widely used word lists is Thorndike and Lorge (1944); however, it is based primarily on adult written text, much of which is no longer in common use. While Kučera and Francis (1967) is a more recent and representative sample of written text, it is clearly a list of words used by adults. Two recent and comprehensive studies of words in elementary textbooks (Harris & Jacobson, 1972; Carroll, Davies, & Richman, 1971) describe what children are expected to be able to read, but not necessarily what they themselves can use.[1] In addition, Carroll et al. covers grades 3-9 rather than K-6. While Rinsland (1945) is a study of children's written words and is therefore a good estimate of children's vocabularies, its usefulness is restricted because of its age.

In addition to the limitations of these lexical studies as sources of words relevant to K-6 instruction, none of them includes the range of linguistic information that accompanies each term in the communication skills lexicon. To help overcome the deficiencies found in individual lexical studies, this compilation is derived from a variety of sources.

The lexicon was originally differentiated into three major components: an entry lexicon, a general lexicon, and a technical lexicon. The entry and general lexicons have since been combined into one compilation, but codes have been included to identify the two

---

[1] These textbook studies were published after most work on compiling the present lexicon was nearly complete; examination of these two sources indicates that their use would not add significantly to this lexicon.

components. The entry lexicon comprises the set of words used by children in kindergarten. While more words are undoubtedly known than are used, word use is more readily studied and has been employed to estimate words known. Although it is probable that a sample of word use cannot be extended to the whole population of kindergarten children, the entry lexicon should at least be representative of words known by children at this level. A set of proper names is also included in this component.

The general lexicon includes those words learned by children after kindergarten and before the end of sixth grade. These words represent contributions of the community and the school to the general vocabulary growth of elementary school children. A set of commonly used geographic names is also included in the general lexicon.

Although not described in this report, the technical lexicon by contrast is the explicit contribution of the school to children's vocabulary; it is the technical terminology introduced by the school in order to teach various school subjects. The technical lexicon might be defined as that set of words commonly used in a discipline, but not commonly used in nonacademic activities, at least not without changes in meaning.

Section I

COMPILATION

The communication skills lexicon has been compiled from a variety of studies of children's speaking, writing, and reading vocabularies. Specific guidelines were applied to delimit the inclusions of words from these sources. This section defines the forms of entries and describes the sources from which they were taken for both the entry and general components of the lexicon.

THE FORM OF LEXICON ENTRIES

Words are included only if they are found as main or run-on entries printed in boldface type in Webster's Seventh New Collegiate Dictionary (1967, hereafter referred to as Webster's) or in The American Heritage Dictionary of the English Language (1969, the AHD).[2] These dictionaries were used because they are comprehensive enough for general use but not so exhaustive as to be beyond the needs of most readers, including elementary school children. The form given in the dictionary is used, especially for determining the status of compounds: as one word or hyphenated (included), or as two or more words unconnected orthographically (excluded). Archaic, obsolete, and objectionable words have been deleted. Inflected forms are included only when they meet the conditions specified below.

Lexicon entries are base forms; i.e., forms without the inflectional suffixes -(e)s (plural nouns or third person singular, present tense verbs), -'s (progressive nouns), -(e)d (past tense or past participle verbs), -ing (progressive verbs forms), -er (comparative adjectives), or -est (superlative adjectives). When the lexicon was being compiled from various sources, words with such inflections were treated as their base forms; it was assumed that if inflected forms are used, the base forms are also known. Inflectional suffixes have been discounted since they apply to most words of the appropriate part of speech without changing the part of speech classification; e.g., fast, faster, and fastest are all adjectives. They are often used automatically as a result of grammatical or semantic context; e.g., in the context "He is talk__" only "ing" is appropriate; in the context "Yesterday he talk__" only "ed" is appropriate. Moreover, by the time of kindergarten entry, most children can make automatic and accurate use of most inflectional suffixes (Berko, 1958).

Although the lexicon is comprised primarily of base forms, inflected forms are included under the following conditions:

---

[2] Words not in these dictionaries are included if they are new words (e.g., tennies, see Appendix A) or words commonly used by children (e.g., ow).

a. The form is irregular; e.g., feet (plural), ate (past tense), eaten (past participle), better (comparative), worst (superlative).

b. The addition of an inflection causes a semantic change; e.g., the plural forms (eye) glasses and (military) arms are semantically different from glass and arm.

c. The word is a plural for which there is no singular form (e.g., scissors, trousers); where the dictionary indicates it is usually plural (e.g., suspenders); or where the noninflected form is not a noun (e.g., the noninflected form of civics is the adjective civic, the noninflected form of pants is the verb pant).

d. The words are -ing and -ed forms which have particular uses as adjectives or (in the case of -ing) as nouns; this was determined as follows:

   (1) A separate main entry in Webster's or the AHD; e.g., dressing, trying, closed, painted.

   (2) Adjective status, as determined by appropriateness in the frame "It is/They are very _____" (cf. Francis, 1958, pp. 268-270); e.g., thrilling, bored.

   (3) Noun status, as determined by appropriateness in the frame "What it needs/they need is a (good) _____;" e.g., cleaning, polishing.

   (4) A base form not commonly used as a verb, particularly by children; e.g., barbed, outing.

   (5) No unsuffixed form; e.g., outstanding.

If, when compiling the lexicon, only an irregularly inflected form was found, the base form was added; e.g., only lice was found, and louse was added.

The most commonly used alternative spellings of words are included; e.g., gray and grey. Some of the sources for the lexicon were based on speech, and speech does not (when alternatives are possible) determine the spelling of a word. For instance, if a source listed dollie, it is entered as both dollie and dolly. This applies also to names; e.g., Allan and Allen. British spellings are not included.

Compounds are included only when written in a single unit or in a hyphenated form, as determined from Webster's or the AHD. (This condition was waived for geographic names so that terms such as United States might be included in the body of the lexicon.) Because of the

high volume of hyphenated fractions and numbers, such words have been excluded; e.g., one-eighth and twenty-two.

Also not represented are localisms and trade names. Only a few abbreviations, very common in speech, are included; e.g., TV. The unique abbreviations Mr., Mrs., and Ms. are included in their graphic forms in order to represent the speech forms /mɪstər/, /mɪsəz/, and /mɪz/.[3]

When the term "words" is used in the following sections of this paper it will mean words and word forms as specified above.

SOURCES FOR THE ENTRY LEXICON

The base source (see Table 1) for the entry lexicon was Kolson (1960), who recorded kindergarten children's speech at home, at school, and in a structured setting where the children responded to pictures. A set of 897,973 running words was collected in the Pittsburgh, Pennsylvania, Washington, D.C., and Portland, Oregon areas, from 494 children chosen by socioeconomic class in proportion to the general population and including 12% nonwhite. This is the most recent and most comprehensive study of words used by kindergarten children. Kolson includes only words that occur at least seven times each. All words from this study were listed for the entry lexicon.

Murphy and others (1957) includes results from several studies of kindergarten children's speech, primarily in New England. Fourteen classrooms were sampled at various times of the year using different methods and yielding 371,281 running words. The study contains all kindergarten words, regardless of frequency. Words with a frequency of at least five were listed for the entry lexicon.

Weaver's (1955) study of kindergarten word use in rural northwest Los Angeles County was used for geographic balance. 121,402 words were collected during regular class activities at various times throughout the year. Words occurring in at least three of the four classes for a total frequency of at least five or in two classes for a total frequency of ten were listed for the entry lexicon. (The more stringent requirements for inclusion were designed to offset the restricted locale of the study.)

Entwisle (1966) conducted a study of 200 kindergarten children in an urban area, Baltimore, Maryland. Over 12,000 word-association responses to 96 stimuli were collected. Since young children give a variety of responses, a considerable number of different words can be found in the study. Those words occurring at least twice as responses to a stimulus were listed for the entry lexicon.

---

[3]When listing geographic terms for computer processing (see Section II), Island was occasionally abbreviated to Isl because of space limitations.

Table 1

Summary of Lexicon Sources

Source	Inclusion Criteria
Entry lexicon	
Kolson (1960	all words
Murphy and others (1957)	kindergarten words with a frequency of 5 or more
Weaver (1955)	words occurring in at least 3 classes with a total frequency of 5 or more, or in 2 classes with a total frequency of 10 or more
Entwisle (1966)	all kindergarten words occurring as responses at least twice to a stimulus
General lexicon	
Entry lexicon	all words occurring only once
Rinsland (1945)	words with a total cumulative frequency of 4 or more, with at least one occurrence at two different grade levels or words with a frequency of 4 or more in grade 8 or words with a frequency of 11 or more at only one grade level, 1-7
Murphy and others (1957)	words with a total cumulative frequency of 5 or more, grades K-3
Jacobs (1967)	all words
Durr (1970)	all words
Green et al. (1958)--Grade 5	all words
Green et al. (1958)--Grade 6	all words
New word study	all words (see Appendix A)

After entry lexicon words had been compiled from all four sources (Kolson, 1960; Murphy and others, 1957; Weaver, 1955; Entwisle, 1966), those words occurring in at least two sources were listed as final entries. All words occurring in only one source were transferred to compilations for the general lexicon (see below).[4]

In addition, the Cronnell (1971b) study of proper names was replicated for first names, including nicknames. On the basis of the changing popularity and frequency of names, a few deletions and additions were made. All names obtained from this replication are included. Second, to provide a more representative sampling of surnames, the 25 highest frequency surnames in Social Security records are included (from The New York Times Encyclopedic Almanac 1972, hereafter referred to as the Almanac). Because of the increasing use of alternate spellings, shortened forms, and stylized spellings, selection of a working list was more subjective than was desirable. Unfortunately, there are no reliable frequency lists of current names that are familiar to children.

SOURCES FOR THE GENERAL LEXICON

Eight sources of words were used to obtain the general lexicon (see Table 1). First, words occurring in only one source for the entry lexicon, therefore not qualifying for inclusion in that first component, were listed for the general lexicon.

Rinsland (1945) contains words written by elementary school children, grades 1-8. (Some of the grade 1 words are from oral sources.) While somewhat outdated, it remains the most comprehensive study of children's vocabulary. Over 100,000 children in 708 different schools contributed over six million running words. Because of its comprehensiveness, it was used as the basic source of the general lexicon. Although the present communication skills lexicon goes only through grade 6, Rinsland's wider grade coverage is acceptable because it indicates productive vocabulary and thus probably underestimates receptive vocabulary. A word from Rinsland was listed for the general lexicon if there was a total frequency of four or more, with at least one occurrence at two different

[4]Many of the closed sets of words in everyday use are known by children in kindergarten. Included from two or more sources were complete sets of terms for relatives, days of the week, seasons, and primary and secondary colors. The occasional absence of words from otherwise complete sets was considered to be coincidental and probably due to sampling effects. Therefore, if at least three-fourths of the words in a defined closed set were found in two or more sources and the remaining words were found in only one source, then all words in the set were included in the entry lexicon. This permitted the inclusion of the following closed sets of terms: money names, months of the year, and number names. All possible closed sets were not defined.

grade levels. Because of the ceiling effect, a word was listed for the general lexicon if there was a total frequency of at least four, all in grade 8. To allow for recognition of terms with relatively high frequencies all in one grade level, a word was listed if there was a total frequency of 11 or more, occurring at only one grade level, 1-7.

In addition, several other sources contributed words to the general lexicon. Murphy and others (1957) contains not only kindergarteners' speech but also words used by first, second, and third graders (19, 22, and 15 classrooms, respectively), primarily in New England. The Murphy study eliminated first, second, and third grade words that occurred fewer than five times. Words from this source with a total cumulative frequency of 5 in grades K-3 (not all in kindergarten) were listed for the general lexicon.

Jacobs (1967) studied free association of over 8,000 students, second through sixth grades, in six Oregon school districts. Subjects wrote down all the words they could think of in a short time. 849,157 running words were obtained and the following were eliminated by Jacobs: words found in the International Kindergarten Union study (Horn et al., 1928) and words occurring fewer than three times. All of Jacobs' words are listed for the general lexicon.

Durr (1970) studied the words used in 80 books commonly selected by primary children for free reading.[5] Out of 105,280 running words, he found 3,220 different words "omitting proper names and onomatopoeic words and counting only the base words of common inflected forms or compound words" (p. 6). No frequency figures are given, but as free-reading choices, the words are clearly read by large numbers of children. In addition, the Durr study is a good source of slang or nonliterary words, sounds, children's words, western words, and animal names. All words from this study were listed for the general lexicon.

Green, Howard, Joerger, and Marino (1958) compiled words used in 20 fifth-grade and 20 sixth-grade textbooks: readers, social studies texts, and science texts. The finished lists (one for each of the two grades) include base forms of words used in two or more texts at the same grade level, excluding words found in Gates (1935). All words from Green et al. were listed for the general lexicon; this provided two sources, one for grade 5 and one for grade 6.

Words occurring in two or more of these seven sources are included in the general lexicon.

Additionally, an eighth source, which made the lexicon more contemporary, was a staff study of new words (see Appendix A). All words

---

[5]The study was done for Houghton Mifflin, from whom the word list was obtained.

from this study are included in the lexicon even if they were not listed from any of the seven other sources.

Other words from only one source are included if they met at least one of the following criteria:

a. Children's words; e.g., beep.

b. Suffixed words for which the base form was in the entry or general lexicon; appropriate suffixes are -(e)s, -ed, -en, -ing, -er/-or, and -ly; e.g., opener.

c. Words suffixed with -er/-or for which only the inflected form was in the entry or general lexicon because the base form is no longer used; e.g., bookkeeping, not bookkeep, was in the lexicon, and bookkeeper from only one source was added.

d. Irregular forms for which the base form was in the entry or general lexicon; e.g., misunderstood.

e. Words in Kučera and Francis (1967) with a frequency listing of 10-05-010 or higher (summed across derived forms if necessary); i.e., occurring at least ten times (per million running words) in at least five genres and in at least ten samples; e.g., function.

f. Words occurring in Thorndike and Lorge (1944) with a frequency of at least ten occurrences per million running words; e.g., overthrow.

Finally, a set of representative geographic (and astronomical) names is included in the general lexicon. These names were determined on the basis of their importance. The source was the Almanac. The categories of geographic names are given in Table 2. Where the criterion is indicated as "selective," the Almanac listings were edited to remove those names believed to be generally unfamiliar to children. "Selective" means that staff members reviewed all items in the Almanac in each category indicated, making subjective judgments about those terms that were important enough to be in general use. Their lists were compared and items of disagreement reconsidered. After further deletions had been made, the combined list of terms in the "selective" categories was edited independently by seven other SWRL staff members selected for their knowledge of children's vocabulary. Those terms agreed upon by four or more of the reviewers are included as the "selective" geographical terms in the general lexicon.

Most of the sources contributing to the lexicon did not include names of people, either real or fictional. Therefore, the lexicon does not contain any of these names.

Table 2

Inclusion Criteria for Geographic Names

A. Political areas	
1. countries	selective
2. states	all U.S. states
3. U.S. territories	selective
4. provinces	all Canadian provinces
5. cities	U.S.: all state capitals
	U.S.: all "Representative American cities" in the Almanac
	Canada: selective
	other parts of the world: selective
B. Physical areas and features	
1. planets	all
2. continents	all
3. islands	selective
4. oceans, seas	selective
5. lakes	selective
6. rivers	selective
7. mountains	selective
8. deserts	selective

Note: The source for all lists was The New York Times Encyclopedic Almanac 1972.

EMPIRICAL VERIFICATION STUDY

After the compilation had been completed, an empirical study was conducted to determine the degree of confidence with which the general lexicon is representative of elementary level children's vocabulary today. Fifteen hundred students in grades 1-6 were tested for knowledge of a sample of 10% of the general lexicon. The results indicated that 95% of the words tested were known[6] by elementary level children, grades 1-6; only 5% of the words tested were known by fewer than half of the children at any grade level.

---

[6]A word was considered "known" if more than half the students at a grade level responded correctly.

Section II

CODING

This section describes the codes used with the lexicon to provide linguistic information. The coding system derives from earlier lexicon work (Berdiansky et al., 1969), but has been revised in the following major ways:

a. The addition of codes to cover more adequately words of more than two syllables, which were not coded in earlier work.

b. The addition of codes for sound-to-spelling correspondences which are not easily retrievable on the basis of spelling-to-sound correspondences (cf. Cronnell, 1971c).

c. The addition of codes for names and for geographical items.

d. The addition of codes for word usage frequency.

e. Revision of codes which were found inadequate in work with the previous lexicon.

A summary of the code types, along with their arrangement for computer entry, is provided in Table 3. The remainder of this section describes in detail the coding system used for the communication skills lexicon.

LINE 1 CODES

Line 1, Field 1: Term

The lexicon term is entered in conventional spelling with internal punctuation where needed. Punctuation is rare: some hyphens in a few compounds, spaces in two-word names, apostrophes in contractions, periods in a limited set of abbreviations. As all letters are upper case, capitalization is ignored (but see Line 1, Field 9).

Line 1, Field 2: Homograph marker

Some orthographic words, while having one spelling, may have two pronunciations corresponding to different meanings or parts of speech; e.g., read, convict. The item is then entered twice, once for each pronunciation, and each entry is coded "1" or "2", the former for the

---

Paula Russell made significant contributions to the design of the codes described in this section.

Table 3

Computer Arrangement of Codes

Field No.	Field Description	Column
LINE 1		
1	Term	1-18
2	Homograph marker	19
3	Line number	20
4	Lexicon component	22
5	Frequency	26
6	Grade level	30
7	Stress and word structure	34-40
8	Compounds	44-48
9	Miscellaneous	50-60
10	Names	62-68
11	Affixation processes	70-80
LINE 2		
1	Term	1-18
2	Homograph marker	19
3	Line number	20
4	Correspondences, affixes, and epenthetic elements	22-60
5	Bases	62-80

most common or most basic use.[6] When a variant pronunciation does not indicate a semantic/syntactic change, only one entry is given (but see Line 1, Field 9) - e.g., either as /ˈiðər/ or /ˈáyðər/, gone as /gɔn/ or /gan/.[7]

Line 1, Field 3: Line marker

The first line for every item is coded "1".

Line 1, Field 4: Lexicon component

Words in the entry lexicon are coded "E"; words in the general lexicon are coded "G".[8]

Line 1, Field 5: Frequency

Frequency is based on Carroll et al. (1971, hereafter CDR). The frequency measure used was the standard frequency index (SFI): the logarithmic transformation of the estimate of true frequency in a theoretically infinite corpus. The SFI is sensitive to the distribution of words across subject areas; e.g., given two words with equal raw frequencies, the one found only in a few subject areas will have a lower SFI than the one found in many subject areas. Based on the SFI, coding is as follows (where "1" is low frequency and "7" is high frequency):

SFI	Lexicon Code
0.0 - 19.9[9]	1
20.0 - 29.9	2
30.0 - 39.9	3
40.0 - 49.9	4
50.0 - 59.9	5
60.0 - 69.9	6
70.0 - 89.9[9]	7

[6]Codes discussed within the text are indicated by quotation marks.

[7]Symbols used in the text are exemplified in Appendix B.

[8]This field is redundant, since all and only "E" words are coded "0" for grade level (Line 1, Field 6). However, should the technical lexicon be incorporated in this list, a lexicon component field would be necessary.

[9]The first two and last two decades are combined because the first and last decades have fewer than 20 words each.

Words occurring only once in CDR are not listed with their SFI's. Instead, they are assigned letters (based on the distribution of items across subject areas), for which the lexicon codes (based on the SFI) are as follows (derived from CDR Table D-1):

CDR	Lexicon Code
C, G, J, K, L, M, S	1
All others (A, B, D, E, F, H, N, P, Q, R)	2

Exact forms were matched (e.g., if uncapitalized in the lexicon, then uncapitalized in CDR). However, if a lexicon base form was not found in CDR, the SFI for its inflected form was used. Inflected forms were used in the following order of preference: -s, -ed, -ing, -'s, -er, -est. If a form inflected in the lexicon was not found inflected in CDR, the SFI for its base form was used. If no other form of a lexicon word was found, then the SFI for its capitalized form was used.

If no form of the lexicon word was found in CDR, then the word was coded "U" (= "unlisted").

Line 1, Field 6: Grade level

Items from the entry lexicon are expected to be known by kindergarten children and were thus coded "0" (=grade one minus one year).[10] Determining grade level for the general lexicon items was much more complex. First, grade levels were obtained from the original sources. Then the grade levels from the sources in which a word was found were combined and averaged.

Grade leveling within sources. Criteria were applied to assign grade levels to words as they were listed from each source (see Table 4).

Lexicon grade leveling. Upon completion of grade level determination for each of the sources in which an item was found, the following leveling rules were applied to obtain the grade level for the lexicon.

a. Grade levels were averaged for words found in the entry lexicon, in Rinsland, and in Murphy and others, or in any combination of two of these three primary sources. All other possible sources and their grade levels were ignored.

[10] The sets of proper names and geographic names (see Section I) were coded at kindergarten ("0") and grade 1 levels respectively.

Table 4

Grade Leveling Criteria

Source	Leveling Criteria
Entry lexicon	grade 1, for items occurring only once
Rinsland (1945)	grade level at which there is the first occurrence of 3; if none, first occurrence of 2; if none, last occurrence of 1 or grade level at which there is a cumulative frequency of 5 if that level is lower than the one obtained above
Murphy and others (1957)	same as "a" above or grade 1, if item occurs in kindergarten and one or more higher grades, and the highest frequency is in kindergarten
Jacobs (1967)	grade level that is listed
Durr (1970)	ungraded
Green et al. (1958)-Grade 5	grade 5
Green et al. (1958)-Grade 6	grade 6
New word study	ungraded

b. Grade levels were averaged for words found in only one of the above three primary sources and also in Jacobs, but the level was never higher than Jacobs. All other possible sources and their grade levels were ignored.

c. The grade level of either the entry lexicon, Rinsland, Murphy and others, or Jacobs was used for words found in any one of these plus one or more of the other four sources. When Rinsland alone was the determinant and the level was 7 or 8, the final grade level was rounded to 6.

d. Grade 5 was used for words found in Green et al.-5, plus Green, et al.-6, and/or Durr and/or the new word study.

e. Grade 6 was used for words found only in Green et al.-6, plus Durr and/or in the new word study.

f. Grade levels were coded "U" (="ungraded") for words found only in Durr and/or in the new word study.

g. Words included from one source only were leveled according to the level of the source, if one was available.

Grade level averages that resulted in whole numbers plus fractions of 1/2 or less were rounded to the lower numbers.

The empirical study conducted on the general lexicon (see Section I) also assessed the degree of reliability of these grade level assignments. The results indicated that 66% of the words tested were known[11] at even lower grade levels than the lexicon indicates--lower by an average of about 2 1/2 grades; 21% were known at exactly the same grade levels specified; 8% were not known (by more than half the children in a grade) until higher levels than the ones specified--higher by an average of about one grade; 5% were known by fewer than half of the children at any grade level.

Grade level assignments in the lexicon are thus conservative, with 87% of the words tested known by a majority of the children by or before the grade level indicated.

Line 1, Field 7: Stress and word structure

Field 7 contains seven columns. In the first column, the number of syllables for the term is indicated; e.g., house = "1", happy = "2", flamingo = "3", automobile = "4", etc. In the second column, the stress pattern is indicated, as follows (ˈ = primary stress; ˋ = secondary

[11] A word was considered "known" if more than half the students at a grade level responded correctly.

stress; - = unstressed); a word ending in a vowel sound and coded with an "A" may have some degree of stress on the final syllable.

one-syllable words: blank

two-syllable words:

A	’-	happy
B	’`	nonsense, convict (noun)
C	-’	above, convict (verb)
D	`’	reread

three-syllable words:

A	’--	happiness, alternate (adj.)
B	’-`	photograph, alternate (verb)
C	-’-	vanilla
D	`-’	engineer
E	’`-	frostbitten
F	`’-	formation
0	other stress patterns	

four-syllable words:

A	-’--	unfortunate
B	-’-`	communicate
C	’-`-	monastery
D	`-’-	automatic
E	’---	gradually
0	other stress patterns	

five-syllable words:

A	-’--- or -’-`-	unfortunately, unnecessary
D	`-’--	hippopotamus
F	-`-’-	congratulation
0	other stress patterns	

words of six or more syllables: all coded "0".

The third column indicates the presence of a compound, pseudocompound, contraction, or abbreviation:

C compound:

a. The word is composed of two or more clearly identifiable free base words without pronunciation or spelling changes; e.g., fireplace.

b. The term is a two- or three-word geographic term (except those that are "G" compounds); e.g., North Dakota.

G pseudocompound:

a. The word includes one or more free bases, but the rest of the word is either a bound base or is not an identifiable base or affix; e.g., microwave, mulberry.

b. The word is composed of two or more free base words, for at least one of which the spelling and/or punctuation differs from the free form; e.g., Christmas, vineyard.

c. The word is a reduplicative compound; e.g., bowwow.

A contraction; e.g., can't, o'clock.

X abbreviation; e.g., Mrs., TV.

If there is a code in this column, there may also be one in Field 8.

The fourth through seventh columns are used to indicate: (4) number of prefixes attached to free bases, (5) number of suffixes attached to free bases, (6) number of prefixes attached to bound bases, and (7) number of suffixes attached to bound bases. These numbers directly reflect codes in Line 2, Field 4.

Line 1, Field 8: Compounds, pseudocompounds, and contractions

When coded "C", "G", or "A" in Line 1, Field 7, third column, the following codes are used where applicable:

H hyphenated compound or pseudocompound; e.g., drip-dry, jack-o-lantern.

S pseudocompound or contraction with spelling change; e.g., cocklebur (from cockle and burr), won't (from will and not).

P pseudocompound or contraction with pronunciation change; e.g., breakfast, mustn't.

M doubled consonant at boundary of component parts; e.g., cattail.

X abbreviation in part of compound; e.g., Washington, D.C.

Multiple codes for a single item are entered in the order given above and are separated by commas.

Line 1, Field 9: Miscellaneous

This field includes a number of features, most of which relate in some fashion to other codings:

S alternate spelling, with both forms coded; e.g., o - oh, grey - gray.

P alternate pronunciation, with no change of meaning or part of speech,[12] e.g., either with /i/ or /ay/. This code is not used: (1) for a pronunciation alternation which is included in the appropriate spelling-to-sound correspondence (cf. Line 2, Field 4); (2) for patterned dialect variations (e.g., for r-less dialects); or (3) for nonstandard forms (e.g., ask as /æks/).

A alternate stress pattern, with no change in meaning or part of speech.[12] This code may be used with or without an alternation in pronunciation; when with a pronunciation alternation, it is also coded "P": e.g., abdomen as ˈ-- or -ˈ-.

D morpheme and/or syllable division must be known to apply spelling-to-sound correspondence; e.g., shepherd with p and h, not ph (→ /f/); the e in homesick as the final silent e in home, not representing a medial syllable.

C capitalized items; e.g., America, Christmas

C+ items occurring in the lexicon both capitalized and uncapitalized, without a change in pronunciation; e.g., Bill and bill, Eve and eve.[13]

V vowel change for inflection; e.g., sat from sit, flew from fly.

K consonant change for inflection; e.g., wives from wife, sent from send.

B consonant cluster simplification at morpheme boundary; e.g., grandmother pronounced /grǽnmʌ̀ðər/, postcard pronounced /póskàrd/.

Multiple codes for a single item are entered in the order given above and are separated by commas.

---

[12] If there is a change of meaning or part of speech, the word is listed twice; see Line 1, Field 2.

[13] Many common nouns are used as names; e.g., Heather, Main Street, Mirror Lake. These are not coded "C+."

Line 1, Field 10: Names

Names, in addition to being coded "C" or "C+" in Line 1, Field 9, are also coded as follows:

1 first names; e.g., John, Jane.

2 last names (as found in the Almanac); e.g., Johnson, Smith. Names that were listed as both first and last names are coded "1, 2"; e.g., Martin, Thomas.

When a name is coded "1" or "1,2" it is also coded with either or both of the following:

M male; e.g., Peter.

F female; e.g., Nancy. Names that were listed as both male and female are coded "M, F"; e.g., Chris, Pat.

One additional name code is used:

G geographic or astronomical names; e.g., Mars, Paris.

Multiple codes for a single item are entered in the order given above and are separated by commas.

Line 1, Field 11: Affixation processes

When one or more affixes (primarily suffixes) are added to a word, a change in the base may result; these change processes are coded as indicated in Table 5. Multiple codes for a single item are entered in the (alphabetical) order given in the table and are separated by commas.

LINE 2 CODES

Line 2, Field 1: Term

Same as Line 1, Field 1.

Line 2, Field 2: Homograph marker

Same as Line 1, Field 2.

Line 2, Field 3: Line number

The second line for every item is coded "2".

Table 5

Codes for Affixation Processes

Code	Process	Example
A	stress and/or pronunciation change; spelling unchanged except as described in codes other than "S", below	resíde → résident
C	ce → t	science → scientist
D	final consonant doubling following a single vowel	rob → robber
DE	d(e) → s	decide → decisive, decision
E	e → Ø	carve → carving
H	hyphen between prefix and base	release → re-release
K	y → ic	apply → application
L	le → il or ul	able → ability; circle → circular
LE	le → Ø before ly	terrible + ly → terribly
LP	/əl/ → /l/	kindle (/kíndəl/) → kindling (/kíndlɪŋ/)
M	doubled consonant at a prefix-base boundary	mis + spell → misspell
O	y → Ø	memory → memorable
R	Vr → r	wonder → wondrous
RP	/ər/ → /r/	flavor (/flévər/) → flavoring (/flévrɪŋ/)
S	spelling change in base not covered by other process codes	permit → permission
T	t → c	prophet → prophecy
V	vowel before vowel-initial suffix → Ø	Mexico → Mexican
W	doubled consonant at base-suffix boundary (one letter from base and other from suffix, as opposed to added letter in "D" code)	mean + ness → meanness
Y	y → i	happy → happiness

Note: Only codes "H" and "M" apply to prefixation; all others apply only to suffixation.

Line 2, Field 4: Correspondences, affixes, and epenthetic elements

All codes in this field are listed consecutively according to the spelling of the item and are separated by commas, with the exception of the following, which are followed by hyphens:

a. Epenthetic elements, except in Greek or Latin compounds.

b. Palatalized and spirantized consonants preceding suffixes.

Correspondences. The basic codes in this field are spelling-to-sound correspondences that are applied to all grapheme units in free bases and in bound bases. These correspondences are derived from those that were originally used in Berdiansky et al. (1969), and that were revised in light of later experience (see Cronnell, 1971a; Berdiansky, Stanton, & Cronnell, 1973). The revised spelling-to-sound correspondences, as used in coding the lexicon, are listed and exemplified in Appendix C; the listings are separate for consonants, secondary vowels (vowel spellings with two or more letters), and primary vowels (vowel spellings with one letter). Besides the addition of new correspondences, the major revision was to change the numbering system for consonants and secondary vowels in order to simplify the coding process. If there is only one correspondence or only one major correspondence, no number is used. If there are two or more correspondences, the major correspondences are numbered 1-4, correspondences for which there is no sound are numbered 5, and less regular or less frequent correspondences are numbered 6-9. The numbering system for primary vowels is, in general, unchanged from earlier versions.

Instances of grapheme units for which there is no productive correspondence (i.e., units having exceptional or irregular spellings) are coded as the grapheme unit, followed by "4" plus another number. (This is similar to the numbering system used in Berdiansky et al., 1969.) The second digit in the code for these exceptions varies as a function of the grapheme unit and the pronunciation, as listed in Appendix D. Abbreviations are coded "#A#".

Additional symbols ("+", "#", and "@") are placed after spelling-to-sound correspondence codes to delimit subsets which are needed for sound-to-spelling correspondences. Moreover, these symbols are used to mark phonetically interesting, but phonologically nonsignificant features in specific items; e.g., vowels before /r/, when r is not part of the defining environment. Uses of these symbols are described in Appendix E.

Affixes. Affixes (prefixes and suffixes) are coded as their basic spelled forms, with numbers added to indicate variations in pronunciation or spelling. An affix added to a free base is followed by "+". Semantically distinct forms of the same graphic affix, as well as affixes that are spelled the same as other codes, are marked "#" or "@" to distinguish them.

Prefix codes are listed in Appendix F. Numeric codes for prefixes are assigned as follows:

No Code: a. The prefix has only one pronunciation; e.g., unlucky.

b. The prefix has several pronunciations:

(1) Stress is on the prefix; e.g., advertise.[14]

(2) The prefix vowel is tense; e.g., prehistoric.

(3) In a two-syllable prefix, stress is on the first syllable; e.g., interfere.

1: a. The pronunciation of the prefix vowel is lax; e.g., preference.

b. In a two-syllable prefix, stress is on the second syllable; e.g., interrogate.

2: a. The prefix vowel is unstressed and may vary in pronunciation and/or in the degree of stress; e.g., precede.

b. In a two-syllable prefix, the vowel in the second syllable is not pronounced; e.g., interest.

3: The pronunciation of the final consonant of the prefix changes but is not reflected in the orthography; e.g., the con in conquer pronounced /kaŋ/.

9: a. The spelling and pronunciation of the final consonant of the prefix are altered to become more like (or exactly like) the first consonant of the base. This is assimilation; e.g., in in insect is coded "IN"; in spelled im in impulse is coded "IN9".

b. Some other spelling change occurs in the prefix; e.g., anti spelled ant in antonym.

The code "PR" is used for any infrequently occurring, unlisted prefix which precedes a free base or a base listed in Appendix H. As with other affix codes, a "+" is added to "PR" if the base is free; e,g, mega in megaphone is coded "PR+", bene in benefit is coded "PR".

Suffix codes are listed in Appendix G. Numeric pronunciation codes for suffixes are assigned in order of frequency and/or expectedness;

[14]The following prefixes occur with varying degrees of stress: MIS, SUB, SUR, SYN, TRANS, and UN. However, no additional code is used because the pronunciation varies so little.

variations cannot be generalized across suffixes as they are across prefixes. The "9" code is added to a suffix code which has any spelling or pronunciation different from those specified. These are three main uses of "9":

a. Some suffixes change stress and/or pronunciation when additional suffixation occurs; e.g., pópular → populárity, ángel → angélic crític → críticize.

b. Some suffixes change spelling as well as pronunciation when additional suffixation occurs; e.g., áble → abílity, círcle → círcular, génerous → generósity.

c. Some suffixes vary in spelling according to preceding elements; e.g., grown coded with "EN9", variety coded with "ITY9".

The code "END" is used for any infrequently occurring, unlisted suffix which follows a free base or a base listed in Appendix H. As with other affix codes, a "+" is added to "END" if the base is free; e.g., dom in freedom is coded "END+", ace in terrace is coded "END".

Epenthetic elements. Epenthetic elements recognized in coding are of three varieties:

a. i or u occurring between a base and a suffix, often causing palatalization of the final consonant of the base.

b. at, it, ic, or ul occurring between a base and a suffix.

c. o or i separating the two elements of a Greek or Latin compound.

The first two kinds of epenthetic elements are followed by hyphens; the third kind is followed by a comma. Numeric symbols are assigned to indicate variant pronunciations as follows:

Code	Pronunciation	Example
I-	merges with and palatalizes preceding consonant	partial
I1-	/i/	centennial
I2-	/y/	brilliant
I3-	/i/ and palatalizes preceding consonant	initiate
U-	/u/ and palatalizes preceding consonant	graduate
U2-	/yu/	annual
AT-	/ət/ or /əč/	imperative, signature
AT1-	/ét/ or /éš/	apparatus, aviation
IT-	/ət/ or /əč/	primitive, furniture

Code	Pronunciation	Example
/Tl-	/ít/ or /íš/	political, position
IC-	/ək/ or /əs/	communicate, magnificent
IC1-	/ík/ or /íš/	particular, judicial
UL-	/yəl/ or /əl/ and palatalizes preceding consonant	regulate, congratulate
0,	/ə/	photograph
01,	/á/	thermometer
1,	/ə/	sacrifice
11,	/í/	participate

The code "EP" is used for any infrequently occurring, unlisted epenthetic element such as the ig in navigate and the iv in cultivate.

Line 2, Field 5: Bases

Free bases. When a word includes a free base, it is listed in this field; e.g., "CARE" as the base for careful. When there are two or more free bases, they are listed consecutively according to the spelling and are separated by commas; e.g., "CHALK, BOARD" for chalkboard. When the base word for a word is composed of a free base plus affixes (and epenthetic elements, where found), they are separated by hyphens; e.g., "CARE-FUL" as the base for carefully.

Three free bases have special codes. To distinguish them from the stressed forms, unstressed man, men and land (pronounced /mən/ or /mɪn/ and /lənd/, as in salesman, salesmen, and Greenland) are coded "MAN9", "MEN9", and "LAND9", respectively.

Bound bases. Latin and Greek bound bases occurring frequently in the vocabulary of elementary school children are listed in Appendix H. Whenever any of these listed bases is found in a lexicon item it is coded in this field; e.g., "DICT" as the base for predict.[15] If the bound base occurs within a free base which is coded in this field, it is separated from other word elements by hyphens; e.g., "PRE-DICT" as the base for prediction.

Listed bases do not necessarily completely represent the pronunciation or spelling. The same spelling may represent more than one pronunciation; e.g., "DEC" in both decade (/dɛk/) and decimal (/dɛs/). If a base spelling varies from the alphabetic code (except for final "silent e," which is omitted in all of these bound base codes), a "9" is added to the code; e.g., claim in exclaim is coded "CLAIM", while clam in exclamation is coded "CLAIM9". Consonant doubling and alternation of y and i are not

[15] While a listed base may be derived from two or more different Latin forms with the same spelling, no differentiation is made; e.g., "TERR" from either terrere "to frighten" (as in terror) or from terra "earth" (as in terrace).

denoted by "9" in this field, for these are products of regular suffixation processes and are represented in the affixation processes codes (Line 1, Field 11). Bound bases that are spelled the same as free bases are coded with "#" to differentiate them.

Unlisted (and therefore infrequent) but recognizable Latin and Greek bases are coded "B"; e.g., for the flor in florist.

CODING AND COMPUTER PROCESSING

The volume of data in the lexicon and the multiple codes on each word would have made it very difficult to maintain accuracy with the usual proofing methods. The task of coding independently obviously demands more precision and alertness than is normally required in proofing. Therefore, instead of having the lexicon proofed, two people coded each term. One person coded all terms, ensuring consistency throughout; the second coder, a role filled alternately by four different staff members, served as a check on the first. The coding of the two staff members was compared. When differences occurred, changes or corrections were made, often after conferring with a third party on points in question.

On matters of morphology, coders consulted the AHD and Webster's. On matters of pronunciation, coders consulted Kenyon and Knott (1953),[16] referring to the AHD and Webster's when Kenyon and Knott seemed idiosyncratic or out-of-date. These authorities were generally consulted only when the coder was aware of variability or peculiarity in the pronunciation of a word. However, the influences of dialect and of one's own idiolect are very strong. It is probable that some words with variant pronunciations were not verified in Kenyon and Knott because the coder was unaware of the possible variations.

---

[16] Generally the first pronunciation in Kenyon and Knott was used, but some deviations were made for simplicity of coding:

a. The following contrasts were not distinguished before /r/: /ɪ/-/i/, /æ/-/ɛ/-/e/, /ɔ/-/o/, /ʊ/-/u/. The code that was closest orthographically was used, followed by "#" (see Appendix E). Thus, words with the same vowel-r sound may be coded with correspondences that indicate different phonetic vowel sounds; e.g., pair ("AI#" = /per/), pare ("A11#" = /per/), pear ("EA6#" = /per/).

b. Instead of Kenyon and Knott's frequent /yu/, /u/ was often used if it was found first in AHD or Webster's. In many environments, this /y/ seems to be dropped in much American speech; e.g., new /nu/.

c. If a medial /ə/ can be optionally dropped in speech, it was generally kept in the coding, but "P,A" was added in Line 1, Field 9.

d. No distinction was made among Kenyon and Knott's /a/, /ɑ/, and /ɒ/; all were considered /a/.

After all the coding had been completed, each line of data was keypunched and verified. Computer retrievals were made of questionable codes for further checking. Such checking was done in two kinds of cases:

a. A few new codes were added during the coding process to reflect features not anticipated in advance; items were checked to ensure that they had these codes when appropriate. For example, only four stress patterns (Line 1, Field 7, second column) were originally used for three-syllable words (i.e., "A", "B", "C", "D",); words with other patterns were originally coded "0". During the coding process patterns "E" and "F" were added; after the coding was completed, all "0" codes were retrieved to be sure that none should be changed to "E" or "F".

b. A few codes were "catch-alls" that were included to mark infrequently occurring features; items with these codes were checked to determine whether any particular feature was frequent enough to require a code of its own. For example, words with prefixes or suffixes, but without free bases or listed bound bases (Appendix H) were coded "B" in Line 2, Field 5; checking of these "B" words indicated several common bases, which were then added to the list of bound bases; e.g., "MICRO" in microfilm, microphone, microscope, microwave.

Codes revised or added during or after the coding process have been included in the preceding description of codes and in the appropriate appendices.

Items changed or corrected were keypunched again. Then the complete keypunched data set of the coded lexicon was alphabetized and stored on computer tape. A computer program was used to eliminate duplications and to check for the presence of a line 1 and line 2 for each item. The listing of this data was also roughly proofed by staff members and gross errors (of keypunching or coding) were corrected. The corrections were made directly on the computer tape by use of an update program. A sample page of the coded lexicon data is shown in Figure 1.

From this data base, computer-generated frequencies were obtained for the various codes. In addition, all items were listed by codes in Fields 4 and 5 of Line 2.[17]

---

[17] Computer tapes and printouts of the coded lexicon are available for use in research. For information, write Division of Resource Services, SWRL Educational Research and Development, 4665 Lampson Avenue, Los Alamitos, CA 90720.

```
CONVINCE        1 G   4   6   2C   1
                2 CON2,V,I16,N,C1#,E20                                  B
CONVINCED       1 G   5   6   2C   11                                     E
                2 CON2,V,I16,N,C1#,ED+                                  CONVINCE
COD             1 C   2   1   1
                2 C2,001@
COOK            1 E   5   0   1
                2 C2,002,K
COOKER          1 G   3   6   2A   1
                2 C2,002,K,ER+                                          COOK
COOKIE          1 E   4   0   2A   1                S
                2 C2,002,K,IE+                                          COOK
COOKING         1 G   5   1   2A   1
                2 C2,002,K,ING+                                         COOK
COOKOUT         1 G   1   1   2BC
                2 C2,002,K,OU,T                                         COOK,OUT
COOKY           1 E   4   0   2A   1                S
                2 C2,002,K,Y+                                           COOK
COOL            1 E   5   0   1
                2 C2,001,L
COOLER          1 G   5   2   2A   1
                2 C2,001,L,ER+                                          COOL
COOLING         1 G   4   3   2A   1
                2 C2,001,L,ING+                                         COOL
COON            1 G   3   2   1
                2 C2,001,N
COONSKIN        1 G   3   5   2BC
                2 C2,001,N,S,K,I15,N                                    COON,SKIN
COOP            1 G   4   2   1                  P
                2 C2,C01,P
COOPERATE       1 G   4   4   48 1  1            D
                2 CON9+,029,P,E17#,R,ATE                                OPER-ATE
COOPERATION     1 G   4   4   5F 11 1            D                      E
                2 CON9+,029,P,E17#,R,ATE,T8-ION+                        CON-OPER-ATE
COOPERATIVE     1 G   4   6   5A 11 1            D,P,A                  A,E
                2 CON9+,029,P,E17#,R,ATE2,IVE+                          CON-OPER-ATE
COP             1 E   4   C   1
                2 C2,C15,P
COP-OUT         1 G   U   U   2BC          H
                2 C2,015,P,OU,T                                         COP,OUT
COPE            1 G   4   7   1
                2 C2,011,P,E18
COPING          1 G   3   5   2A   1
                2 C2,013,P,ING
COPPER          1 G   5   3   2A   1
                2 C2,016,PP,ER
COPPERHEAD      1 G   3   2   3BC  1
                2 C2,016,PP,ER,H,EA6,D                                  COPPER,HEAD
COPY            1 E   6   0   2A   1
                2 C2,033,P,Y
```

Figure 1. Sample page from computer printout.

Section III

ANALYSIS

This section reports on the linguistic information obtained through the codes that were applied to the terms. Similar features in each field were examined and some of the resulting data have been analyzed. Frequencies are reported for every field in which such information is significant.[18]

LINE 1 CODES

Line 1, Field 2: Homograph marker

The lexicon includes 96 homograph pairs (192 terms), representing less than 2% of the lexicon. Judgment about children's familiarity with the alternate form of a homograph was subjective. Unless it was felt that both forms were in relatively common use, the word was not coded as a homograph.

One-third of the homographs are pairs of words that are semantically unrelated and that have different pronunciations; e.g., Polish /pólɪš/ (pertaining to Poland) and polish /pálɪš/ (shine), sow /so/ (plant seeds) and sow /saw/ (female hog).

Two-thirds of the homographs are pairs of words that are semantically related, but that represent two different form classes; e.g., close /kloz/ (verb) and close /klos/ (adjective), suspect /səspékt/ (verb) and suspect /sʌ́spɛ̀kt/ (noun). The two words of such pairs have different stress patterns and/or vowel or consonant correspondences. Therefore, they are designated as homographs, and the two forms are coded as separate items.

Line 1, Field 5: Frequency

The frequency measure used was the standard frequency index (SFI) from CDR (see Section II, pp. 15-16). Frequency codes in the lexicon correspond to the SFI decades; e.g., SFI's of 40.0 - 49.9 are coded "4". CDR is not only the most recent word frequency count, but it is also the most extensive. However, the CDR sample includes words in running text taken primarily from textbooks, while the lexicon sample includes words taken from studies of terms that children know and use. Despite

---

[18] The following fields are not reported on in this section:

Lines 1 and 2, Field 1 - Term
Lines 1 and 2, Field 3 - Line number
Line 1, Field 4 - Lexicon component (covered by Line 1, Field 6)

the differences, a comparison was made of low to high frequency words in the lexicon and low to high frequency words in CDR.

Comparison of frequencies in the lexicon and in CDR. Almost half of the 86,741 different "word types" in CDR are low frequency items (SFI's of 01.0 - 19.9 and 20.0 - 29.9), called by CDR "rare word types"; e.g., ethnology, perjury, regenerative. Textbooks introduce many such words that students are expected to learn. Indications are that K-6 children do not, in fact, know many of these "rare word types." Only 2% of the words in the lexicon are low frequency items (codes "1" and "2"); 53% of the "word types" in CDR have these low frequencies (see Figure 2).

Different factors combine to skew the frequency results of both the lexicon and CDR. First, efforts were made to exclude low frequency items from the lexicon and to obtain primarily high and middle frequency words. As the CDR sample was taken from running text, it did not exclude anything.

Second, the lexicon defined a word as a base form that is listed in either Webster's or the AHD, or as an inflected form that satisfies specified conditions. CDR defined a word as a string of graphic characters bounded at left and right by space; graphic characters included letters, numerals, internal punctuation (hyphens and apostrophes), and some mathematical symbols. This resulted in inclusion of such forms as capitalized terms (from titles, initial words of sentences, etc.), all inflected forms, unusual compounds not listed in dictionaries, and both numbers and dates. CDR includes, for example, each of the following as a separate "word type": all, All, ALL, all's, plus 22 hyphenated compounds with all- (e.g., All-the-Cow-there-was, all-water); sometime-editor, gleeman, roundnose; 15-12, 1569-1589, and also 1589.

Third, a comparison of frequency results becomes difficult when there is variance in the grade level coverage of the two studies. The lexicon covers grades K-6; CDR covers grades 3-9. As would be expected, the words in the lexicon that are used by children in grades K-2 did also occur at higher grade levels and were included in CDR. However, most of the low frequency terms found in textbooks for grades 7-9 will not be in the lexicon. These terms increase the percentage of low frequency, "rare word types" in CDR.

In looking at a comparison of the high frequencies (see Figure 2), it should be remembered that high frequency words with SFI's of 70.0 - 89.9 in CDR and codes of "7" in the lexicon are primarily function words (e.g., articles, conjunctions, prepositions, pronouns, verbal auxiliaries, etc.) plus a few very high frequency words such as said, see, made, get. The number of these terms is almost constant and will represent a smaller percentage of an 86,000 word study than of a 10,000 word study.

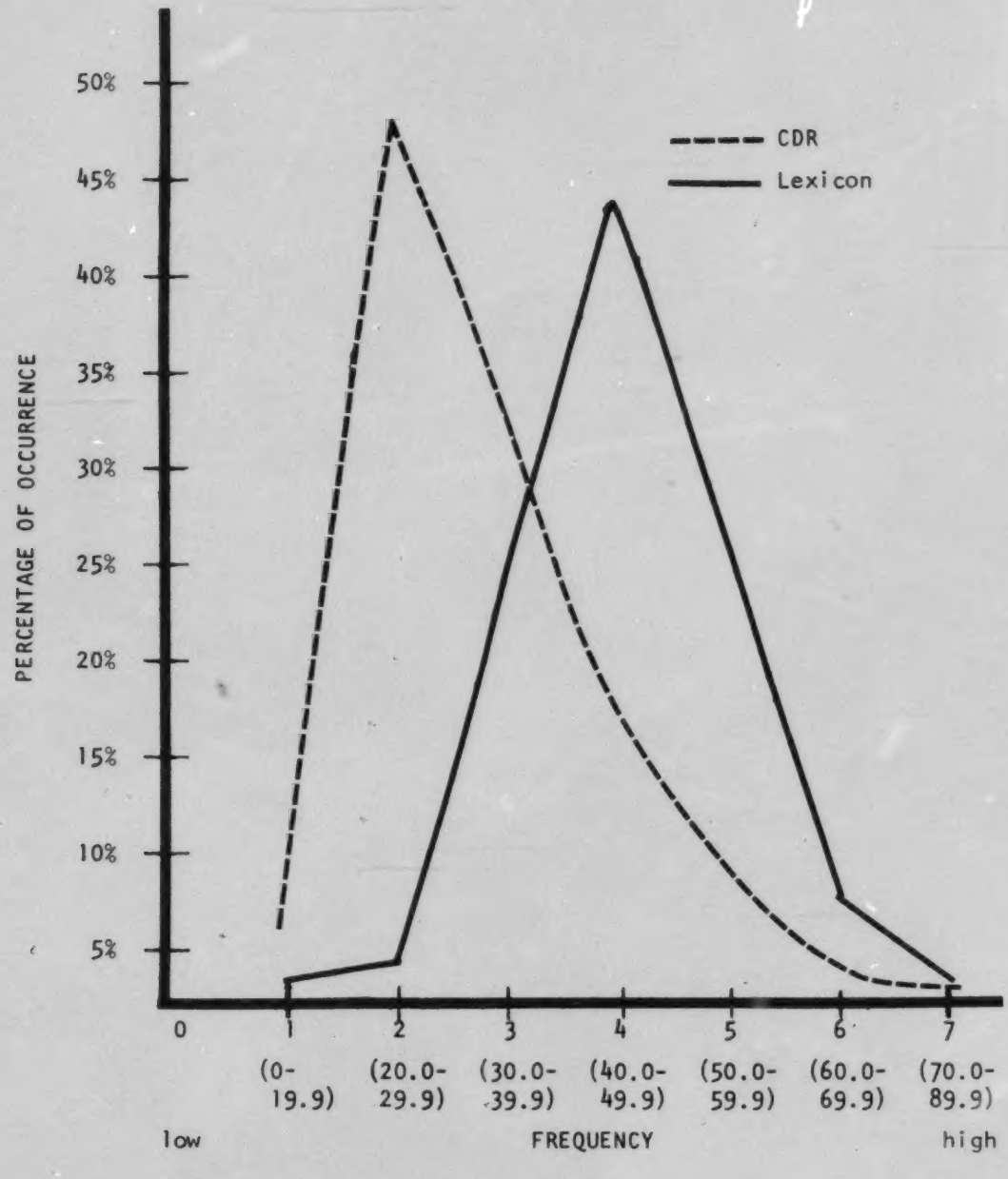

Figure 2. Occurrence of low to high frequency words. Standard Frequency Data (in parentheses) are from Carroll, et al., 1971.

The frequency profile for CDR would be much different if almost 50% of the CDR "word types" were not low frequency items. As it stands, approximately 86% of the lexicon words are mid-frequency items (codes "3", "4", "5"), while about 45% of the "word types" in CDR reflect middle frequencies (SFI's of 30.0 - 59.9).

Lexicon frequencies by grade level. The words with the highest frequencies (codes "7" and "6") are known early in a child's life. Kindergarteners know 100% of the words coded "7", and 80% of the words coded "6" (see Table 6). They even know 40% of the words coded "5", a higher percentage than is represented at any other one grade level. In fact, over two-thirds of the kindergarten-level words have frequency codes of "7", "6", and "5". However, with the large category of words coded "4" and "3", representing 60% of the lexicon, are found in higher percentages in sixth grade than in any other grade. The relatively small number of words with low-frequency codes of "2" and "1" are scattered across the grade levels.

It should be noted that a higher percentage of words in the lexicon occurs at kindergarten level than at any other grade level. This accounts in part for the relatively high percentages of kindergarten-level words at most frequencies.

Line 1, Field 6: Grade level

Approximately one-fourth of the lexicon consists of words known by children at the kindergarten level (see Table 7). This represents the words learned from preschool age through kindergarten. As would be expected, the number of words acquired in first grade is less than half the number acquired before first grade (i.e., "K"). In fact, there would have been even more words at kindergarten level and fewer at grade 1, if the leveling criteria had been different. A word occurring only once in an entry lexicon source, and therefore not qualifying for inclusion in that component, was assigned a grade level of 1 for the general lexicon. This grade level assignment should have been 0, for kindergarten. Then when a word was found in one kindergarten source and in one source at first grade level, the average of the two grades (0 and 1) would have been rounded down to zero. Instead, such a word was leveled at 1.

At each level through grade 4, the number of words added to children's vocabularies continues to decline. However, in grades 5 and 6, there is an increase. This is at least partially attributable to the use of the sources, Green et al., grades 5 and 6, which compiled words from reading, social science, and science textbooks. Another factor in the increase of vocabulary at grade 6 was the leveling procedure for items occurring in another source, Rinsland, at grade levels 7 and 8. If Rinsland alone was the determinant of grade level, the 7 or 8 was rounded to 6.

Table 6

Percentages of High to Low Frequency Words by Grade Level

Frequency Code	Grade Level								Total	
	K	1	2	3	4	5	6	Ungraded	Occurrence of Freq. Codes	% of Lexicon
7	100%	--	--	--	--	--	--	--	108	1%
6	80%	10%	6%	2%	2%	*	--	--	644	6%
5	40%	13%	12%	9%	9%	8%	9%	*	2630	26%
4	14%	12%	11%	12%	10%	12%	29%	1%	4244	42%
3	10%	11%	9%	11%	10%	17%	28%	4%	1866	18%
2	18%	16%	4%	11%	4%	16%	19%	12%	183	2%
1	8%	31%	--	15%	8%	23%	15%	--	13	*
Unlisted	7%	24%	6%	5%	4%	5%	7%	43%	429	4%
Total									10,117	99%

*Less than 0.5%

Table 7

Grade Level Frequencies

Grade Level	Frequency	% of Lexicon
K	2505	25%
1	1212	12%
2	1033	10%
3	985	10%
4	902	9%
5	1126	11%
6	2055	20%
U	299	3%
Total	10,117	100%

The code "U", meaning "ungraded," denotes items obtained in the new word study (see Appendix A), or items from Durr (a study of library books read by children).

Line 1, Field 7: Stress and word structure

Syllables. The first column of this field provides information about the number of syllables in each term (see Table 8). More than one-fourth of the words in the lexicon have only one syllable. However, there are over 50% more two-syllable words than one-syllable words; together they constitute 73% of the lexicon. The remaining terms include 19% with three syllables, 6% with four syllables, slightly over 1% with five syllables, and less than 9.2% with six syllables.

Stress. Although the second column of this field carries a letter code to designate the location of stress in a word, the pattern of stress is defined by the number of syllables. Efforts were made to coordinate similar stress patterns in multisyllabic words by the use of the same code letter; e.g., "2D" (`'), "3D" (`-'), "4D" (`-'-), "5D" (`-'--).

Table 8

Frequencies of Syllables and Stress Patterns

Syllable and Stress Code	Stress Pattern	Stress Pattern Frequency	Syllable Frequency	% of Lexicon
1		2841	2841	28
2A	´-	2893	4525	45
2B	ˆ	809		
2C	-´	670		
2D	`	153		
3A	´--	648	1961	19
3B	´-`	399		
3C	-´-	534		
3D	`-´	100		
3E	ˆ-	141		
3F	`-	129		
30	other	10		
4A	-´--	219	643	6
4B	-´-`	33		
4C	´-`-	105		
4D	`-´-	217		
4E	´---	36		
40	other	33		
5A	-´---	25	130	1
5D	-´-`-	48		
5F	`-´--	26		
50	other	31		
60	other	17	17	*
Totals		10,117	10,117	99

*Less than 0.5%

However, it was not always possible to maintain such a consistency in the codes that were established; e.g., "2C" (-ˊ), "3C" (-ˊ-), "4C" (ˊ-ˋ-).

The pattern found most often in the lexicon is "2A", i.e., two-syllable words having primary stress on the first syllable and no stress on the last syllable; e.g., ápple, háppy. These "2A" words outnumber even the one-syllable words, and comprise more than 28% of the lexicon.

The frequency of occurrence of other stress patterns is much lower. Approximately 8% of the lexicon words have the "2B" stress pattern, i.e., two syllables with primary stress on the first syllable and secondary stress on the last syllable; e.g., búrlàp, cónvìct (as a noun). The "2B" stress pattern is commonly found in compound words; 71% of the words with this pattern are compounds or pseudocompounds; e.g., keyhole, pastime.

The reversal of the frequently occurring "2A" pattern is "2C", i.e., two-syllable words having no stress on the first syllable and primary stress on the last syllable; e.g., distúrb, propóse. The 7% represented in this group are primarily prefixed forms with Latin bases.

The largest group of three-syllable words consist of those with a "3A" pattern, i.e., three-syllable words having primary stress on the first syllable and no stress on either the middle or the last syllable; e.g., cápable, grácìously. These words account for a little more than 6% of the lexicon. Over 77% of the "3A" words have a base in the stressed first syllable, followed by one or more suffixes.

Slightly less than 6% of the lexicon are "3C" words, i.e., three syllables with primary stress on the second syllable and no stress on the first or the last syllable; e.g., dishónest, inféction. Almost three-fourths of these terms have a base in the stressed second syllable, preceded and followed by affixation. In fact, almost half are Latin-base words with a "2C" pattern plus a suffix; e.g., dispóse + al → dispósal, projéct + or → projéctor.

Approximately 4% of the words in the lexicon have the "3B" stress pattern, i.e., three syllables having primary stress on the first syllable, no stress on the second syllable, and secondary stress on the last syllable; e.g., árchitèct, órganìze. Almost half of these terms are compounds or pseudocompounds. Just as two-syllable compound words often have a "2B" stress pattern, three-syllable compound words often have a "3B" stress pattern; e.g., bárbershòp, eárthenwàre.

The total of all the above-mentioned stress patterns comprise 83% of the multisyllabic words in the lexicon. The other 17% are distributed among 12 specific patterns and 4 unspecified potpourri groups designated by the numbers of syllables plus "0" (again see Table 8).

Compounds, pseudocompounds, contractions, and abbreviations. These categories are designated in the third column of this field. The lexicon contains 1,092 compounds and 227 pseudocompounds; together, they comprise 13% of the lexicon (see Table 9).

Table 9

Frequencies of Compounds, Pseudocompounds, Contractions, and Abbreviations

Feature	Code	Frequency		% of Lexicon	
Compounds	C	1092	1319	11%	13%
Pseudocompounds	G	227		2	
Contractions	A	44		*	
Abbreviations	X	5		*	

*Less than 0.5%

The stress patterns that occur with high frequencies in compound words include: 54% with the "2B" (ˆˋ) pattern, 13% with the "3B" (ˊ-ˋ) pattern, 10% with the "3E" (ˆ-) pattern, and 5% with the "2D" (ˋˊ) pattern. The remaining compound words, comprising 18%, are distributed among 16 other multisyllable patterns. (There were no occurrences of compounds with either a "3C" or "5A" pattern.)

Most of the words that are coded as pseudocompounds are composed of two ore more free bases, at least one of which has undergone a spelling and/or pronunciation change; e.g., shepherd, necklace. Those terms in which the pronunciation of at least one base has changed will not have the stress patterns common to compounds. Many terms like cúpboard and cóngressman have completely lost the normal stress on the second base.

The lexicon includes 44 contractions that can be classified into two groups. First, the word not is contracted on all inflected forms of the verbal auxiliaries be, have, and do, and on the modal auxiliaries can, could, will, would, should, and must. The only negative contractions listed in the dictionaries but not occurring in the lexicon sources are mayn't, needn't, oughn't, and shan't. Second, the verbal auxiliaries am, are, is/has[19], had/would[19], have, and will are contracted on complete paradigmatic sets of personal pronouns. All of these contractions were found at the kindergarten level.

[19]The two sets is/has and had/would represent ambiguities; in each set, the members have the same contracted forms, 's and 'd.

Since being listed in Webster's or the AHD was a requirement for inclusion, many commonly used spoken contractions will not be found in the lexicon. These include the above verbal auxiliaries contracted on relative and interrogative pronouns (e.g., that's, who'll) and on adverbials (e.g., how'd, where've).

However, two additional common contractions are in the lexicon; i.e., ma'am and o'clock.

Abbreviations were generally not included; but the following five items were admitted: Mr., Mrs., Ms., OK, and TV.

Affixes. The fourth through seventh columns of Field 7 indicate the numbers of prefixes and suffixes that occur per term (see Table 10).

Table 10

Frequencies of Affix Occurrence

No. of Affixes Per Word	No. of Affixed Words	Total Affix Frequency	% of Total Frequency
1 prefix	1820	1846	23%
2 prefixes	13		
1 suffix	5529	6161	77%
2 suffixes	301		
3 suffixes	6		
4 suffixes	3		
Total		8007	100%

The total frequency of affixes is over 8000, of which 23% are prefixes and 77%, suffixes. One-fourth of the prefixes are attached to free bases (e.g., de + tour → detour, re + fill → refill), while one-half of the suffixes are attached to free bases (e.g., block + ade → blockade, class + ify → classify). Over 99% of the total number of prefixed words have only one prefix, but 13 words have two prefixes (e.g., dis + ad [→ap] + point → disappoint, re + in + force →

reinforce). Almost 95% of the total number of suffixed words have only one suffix, but 5% have two suffixes (e.g., rot + at[e] + ion → rotation, trav + el + er → traveler), 6 words have 3 suffixes, and 3 words have 4 suffixes.

It should be pointed out that there is some overlap of the words in which the affix combinations occur. Not only do one or two prefixes occur in some words, and one, two, three, or four suffixes occur in other words, but also various combinations of both prefixes and suffixes occur in many terms (e.g., con + stit [base] + ut[e] + ion + al → constitutional, dis + ad + vant [base] + ag[e] + ed → disadvantaged, super + in + tend [base] + ent → superintendent). Thus, the total number of different affixed words (almost 6,000) is considerably less than the total number of affixes (over 8,000). Of the almost 6,000 affixed terms, 16% are both prefixed and suffixed. The amount and complexity of affixation in the lexicon is significant.

Line 1, Field 8: Compounds, pseudocompounds, and contractions

This field is coded to provide additional information about some of the compounds, pseudocompounds, and contractions that comprise over 13% of the lexicon (see Table 11). (These categories were designated in Field 7.)

Table 11

Frequencies of Codes About Compounds, Pseudocompounds, and Contractions

Code	Compound	Pseudo-compound	Contraction	Total
H	109	18	---	127
P	---	116	2	118
S	---	46	3	49
M	15	2	---	17
X	---	5	---	5

The code "H" indicates hyphenated forms of compounds and pseudocompounds. Hyphenation occurs in 10% of these two combining forms; e.g., brother-in-law (a compound), X-ray (a pseudocompound).

Changes in pronunciation of the component words, coded "P", are found primarily in pseudocompounds; e.g., gentleman → /ǰéntəlmən/ instead of /ǰéntəlmæn/, Wednesday → /wénzdi/ instead of /wédnəsdè/. Over half of the pseudocompounds have undergone this kind of change. A small number of contractions also have pronunciation changes; e.g., mustn't → /mÁsənt/ instead of /mÁstənt/.

Spelling changes coded "S" are also found primarily in pseudocompounds; e.g., Christmas instead of Christmass, welfare instead of wellfare. Approximately one-fifth of the pseudocompounds have had spelling changes. A few contractions have had more radical spelling changes than are customarily signaled by apostrophes; e.g., o'clock from of + the + clock, won't from will + not.

Approximately 29% of the pseudocompounds did not have spelling or pronunciation changes. These were either (1) reduplicatives; e.g., dillydally; (2) combinations of free bases and bound bases (mostly Greek); e.g., agriculture; or (3) combinations of free bases and unidentifiable bases or affixes; e.g., muskrat.

The code "M" identifies a combination in which there is a doubled consonant at the boundary of component parts; e.g., earring (a compound) bowwow (a reduplicative pseudocompound). This orthographic feature is found in about 1% of the total group of compounds and pseudocompounds.

Terms that have an abbreviation in part of the compound were coded "X". All terms with this code except the unique word T-shirt are geographic phrases; e.g., St. Louis.

Line 1, Field 9: Miscellaneous

Several kinds of miscellaneous information are coded in this field (see Table 12). The largest category consists of words that have alternate pronunciations. Usually these variant pronunciations involve vowel changes. Such words are coded "P". Many represent regional dialect differences; e.g., aunt → /ænt/ and /ant/, creek → /krik/ and /krɪk/, Theresa → /tərísə/ and /tərésə/. Many others indicate idiolect differences with vowel or consonant alternations; e.g., pajamas → /pəǰáməz/ or /pəǰáməz/ with vowel alternation, February → /fébruèri/ and /fébyuèri/ with consonant alternation. Seven percent of the words in the lexicon are coded for alternate pronunciations.

Almost half of the terms coded "P" for an alternate pronunciation are also coded "A" for an alternate stress pattern; e.g., address → /ədrés/ and /ædrès/, adult → /ədÁlt/ and /ædÀlt/. Most of the words with both alternate pronunciation and alternate stress have a possible syllable reduction; e.g., chocolate → /čɔ́kələt/ and /čɔ́klət/, towel → /táwəl/ and /tawl/. Almost four percent of the words in the lexicon are coded for an alternate stress pattern.

Table 12

Frequencies of Miscellaneous Features

Codes	Frequencies	% of Lexicon
P	674	6.6%
A	404	3.9%
S	103	1.0%
C	664	6.5%
C+	53	0.5%
D	385	3.8%
V	178	1.7%
K	66	0.6%
B	72	0.7%

Words having alternate spellings are coded with an "S". About one-fifth of these are in the set of proper names introduced at kindergarten level; e.g., Ann and Anne, Lewis and Louis. Most of the rest of the "S" terms indicate that spelling changes toward simplification are in progress in the language, but have not been completed; e.g., ax and axe, mold and mould, sulfur and sulphur. One percent of the words in the lexicon are coded to mark alternate spellings.

Capitalized terms are coded "C". The large number of these items is primarily the result of the inclusion of a set of 394 proper names at kindergarten level and a set of 241 geographic names at first grade level. Included in the other capitalized forms are names of the months, days of the week, holidays, names of languages, and group of people, etc.; e.g., Chinese, Indian. Terms that can occur in both capitalized and uncapitalized forms are coded "C+". Most are proper names that are also commonly used as uncapitalized words with different meanings; e.g., Art and art, Bill and bill, Ginger and ginger, Pat and pat. The "C" and "C+" forms combined comprise 7% of the lexicon.

In some words, morpheme and/or syllable division must be known to apply spelling-to-sound correspondences. Such terms are coded "D". Almost all of these words are either compounds, pseudocompounds, or affixed forms. The majority of the words have final "silent" e before

an added morpheme; e.g., basement → /bésmənt/, not some pronunciation such as /bésəmènt/, and pineapple → /páynæpəl/, not some pronunciation such as /pɪnǽpəl/. The other words coded with "D" are usually compounds in which the final consonants of the first words could be incorrectly combined with adjacent consonants to form digraphs; e.g., arrowhead with ow and h, not wh → /hw/, lighthouse with t and h, not th → /θ/, uphill with p and h, not ph → /f/. Occasionally, a prefix or suffix attached to a base forms an unrelated orthographic unit; e.g., (bet)(ray), (coin)cidence, nu(clear), (rein)force. Almost 4% of the words in the lexicon are coded with "D".

Those words in which inflection is accompanied by a vowel change in spelling and/or pronunciation are coded "V". These are primarily past tense and past participle forms of verbs that have not been regularized; e.g., bought, froze, sung. A few of the words coded for vowel changes with inflection are plural nouns; e.g., feet, men, mice. Almost 2% of the words in the lexicon are coded with "V".

In some terms, inflection is accompanied by a consonant change in spelling and/or pronunciation. Such items are coded "K". Most words having consonant changes are strong verbs; many of them also have vowel changes, coded "V". Like the "V" items, they are primarily high frequency words. The consonant changes in verbs are often found in irregular past tense and past participle forms; e.g., left, taught. Some of the words coded for consonant changes with inflection are plural nouns; e.g., lice, mice. In a noun, the change more often occurs in the base when the word is suffixed; e.g., leaf + s → leaves, yourself + s → yourselves. Words coded "K" comprise 0.6% of the lexicon.

In a relatively small group of words, there is consonant cluster simplification at morpheme boundaries. Such words are coded "B"; e.g., Christmas → /krísməs/ not /krístməs/, handkerchief → /hǽŋkərčəf/ not /hǽndkərčəf/, Wednesday → /wɛ́nzdi/ not /wɛ́dnəsdi/. Words coded with "B" comprise 0.7% of the lexicon.

Line 1, Field 10: Names

The two kinds of names coded in this field are geographic terms and proper names (see Table 13).

The set of geographic (and astronomical) terms incorporated in the lexicon at first-grade level consists of names of continents, countries, cities, oceans, rivers, etc. Most of these terms, which are coded "G", are one-word names. However, a few geographic terms that are phrases were included in the lexicon; e.g., Gulf of Mexico, Rio de Janeiro, United States. The lexicon contains 241 terms pertaining to geography and astronomy.

Table 13

Frequencies of Names

Feature	Code	Frequency	Total	% of Lexicon
Geographic terms	G		241	2%
First names	1			
Male names	M	188		
Female names	F	197		
Total M + F		385		
Either male or female	M,F	- 13		
Total different first names		372		
Last names	2	25		
Total 1 + 2		397		
Either first or last	1,2	- 3		
Total different proper names			394	4%
Grant Total			635	6%

The set of proper names was included in the lexicon at kindergarten level. These names are designated as first names, coded "1", or last names, coded "2". There are 372 first names and 25 last names, with three of the last names coded as appropriate for either first or last names; i.e., Lewis, Martin, and Thomas. First names are also designated as male names, coded "M", and female names, coded "F". There are 188 male names and 197 female names, with 13 of the total coded as appropriate for either male or female names; e.g., Chris, Jerry. The lexicon contains a total of 394 proper names.

Together, these two sets--geographic terms and proper names--represent slightly more than 6% of the lexicon.

Line 1, Field 11: Affixation processes

Codes in this field designate many of the orthographic features that exist and the spelling and/or pronunciation changes that occur when suffixes and prefixes are added to base forms (see Table 14).

Table 14

Frequencies of Affixation Processes

Code	Frequency	% of Lexicon
A	342	3.4
C	6	0.1
D	156	1.5
DE	15	0.1
E	700	6.9
H	1	*
K	10	0.1
L	21	0.2
LE	8	0.1
LP	33	0.3
M	251	2.5
O	26	0.3
R	27	0.3
RP	33	0.3
S	156	1.5
T	12	0.1
V	12	0.1
W	40	0.4
Y	89	0.9
Total	1938	19.1

*Less than 0.05%

However, the total number of codes here is not relative to the total number of words receiving these codes. Many words have two or three codes for affixation processes; therefore, the number of words involved will be much lower than the number of codes. Also, most items that are affixed do not undergo these spelling and/or pronunciation changes. Approximately 6000 words in the lexicon have been affixed, but only 2000 of these codes have been designated on an even smaller number of words.

Three different pronunciation changes are designated in this field by the codes "A", "LP", and "RP". The code "A" denotes a process in which derivational suffixation causes a stress and/or pronunciation change (especially in a vowel pronunciation) in the base; e.g., accident /ǽksədənt/ + al → accidental /æksədɛ́ntəl/, preside /prizáyd/ + ent → president /prɛ́zədənt/. Both of the examples cited have changes in stress as well as in vowel pronunciation. If the suffixed form is suffixed again, but without another change in stress or pronunciation, the "A" is not coded again; e.g., àccidéntal + ly → àccidéntally. However, if the suffixed form is suffixed again, with another change in stress and/or pronunciation, the "A" is coded again; e.g., president /prɛ́zədənt/ + i-al → presidential /prèzədɛ́nšəl/. "A" represents 18% of the codes in this field, and is designated on over 3% of the words in the lexicon.

The code "LP" denotes a pronunciation change that usually occurs in words with le (/əl/) suffixes. When another suffix is added, the syllable is reduced to /l/, deleting the vowel. Most of the words with this code are verbs to which either er or ing has been added; e.g., sparkle /spárkəl/ + er → sparkler /spárklər/, handle /hándəl/ + ing → handling /hándlɪŋ/. "LP" is designated on 0.3% of the words in the lexicon.

The code "RP" denotes a pronunciation change that often occurs in words that have final Vr. When a suffix is added, the syllable is reduced to /r/, deleting the vowel. Sometimes this vowel deletion occurs in a base (e.g., [dis + fer (base) →] differ /dɪ́fər/ + ent → different /dɪ́frənt/), and sometimes it occurs in a suffix already attached to a base (e.g., [lab + or →] labor /lébər/ + at-ory → laboratory /lǽbrətòri/). "RP" is designated on approximately 0.3% of the words in the lexicon.

Three orthographic features are designated in this field by the codes "H", "M", and "W". The code "H" indicates that a word has a hyphen between a prefix and a base. There was only one occurrence of this in the lexicon; i.e., re-release.

The code "M" is applied when the addition of a prefix results in a doubled consonant at a prefix-base boundary; e.g., dis + satisfy → dissatisfy, un + necessary → unnecessary. Consonant-final prefixes are often altered to make the final consonant more like (or exactly like) the first consonant of the base. This assimilation process occurs most

frequently in words having bound Latin or Greek bases, and not frequently in words having free bases like those cited above. Words in which the final consonant of the prefix has been completely assimilated account for most of the "M" codes; e.g., ad + cord → accord, con + mit → commit. "M" is designated on approximately 2.5% of the words in the lexicon.

The code "W" is applied when the addition of a suffix results in a doubled consonant at a base-suffix boundary. Although this code anticipated forms like heel + less → heelless and mean + ness → meanness, no such words happened to occur in the lexicon. All of the items coded with "W" originally had final l, and were suffixed with ly. Furthermore, all of these l-final words were already suffixed with either al, el, or ful; e.g., fin + al + ly → finally, cru + el + ly → cruelly, peace + ful + ly → peacefully. "W" is designated on 0.4% of the words in the lexicon.

Several kinds of spelling changes, sometimes affecting pronunciation, are also coded in this field. One major category designates processes in which the final letter(s) of a word is deleted with suffixation. The code "E" indicates that a final "silent" e is deleted when a suffix is added; e.g., engine + eer → engineer, noise + y → noisy. "E" represents 36% of the codes in this field, and is designated on 7% of the words in the lexicon.

The code "LE" indicates that a final le is deleted when the suffix ly is added; e.g., comfortable + ly → comfortably, gentle + ly → gently. As evidenced in the examples, the final le can be either the suffix le, or part of a suffix like able that ends in le. "LE" is designated on less than 0.1% of the words in the lexicon.

The code "V" indicates that a final vowel is deleted when a vowel-initial suffix is added; e.g., idea + al → ideal, Mexico + an → Mexican, two + ice → twice. "V" is designated on 0.1% of the words in the lexicon.

Three different types of spelling change occur on y-final words. First, the code "O" indicates that a final y is deleted when another suffix is added; e.g., memory + able → memorable, study + ent → student. "O" is designated on approximately 0.3% of the words in the lexicon.

Second, the code "K" marks a change in which a final y is replaced by ic; e.g., apply + ant → applicant, history + al → historical. "K" is designated on 0.1% of the words in the lexicon.

Third, the code "Y" marks a change from final y to i; e.g., easy + ly → easily, marry + age → marriage. "Y" is designated on 0.9% of the words in the lexicon.

Two processes of spelling change operate conversely and are coded "T" and "C". The code "T" marks a change in which a final t or te

changes to c with suffixation; e.g., benefit + i-al → beneficial, delicate + y → delicacy. "T" is designated on 0.1% of the words in the lexicon.

The code "C" marks a change in which a final ce changes to t with suffixation; romance + ic → romantic, science + ist → scientist. The process coded by "C" does not occur as often as the converse process, "T". "C" is designated on less than 0.1% of the words in the lexicon.

The following kinds of spelling change occur with suffixation, but they are not related. The code "D" indicates that when a suffix is added to a word that has a final consonant preceded by a single vowel letter, the consonant is doubled. (This process differs from that coded "W", in which no letter is added.) The base forms affected by this productive suffixation process are free bases--usually simple, one-syllable words; e.g., fun + y → funny, wed + ing → wedding. A few of the "D" items are two-syllable, prefixed words; e.g., (re + bel →) rebel + ion → rebellion, (trans + mit →) transmit + er → transmitter. "D" is designated on approximately 2% of the words in the lexicon.

The code "DE" marks a spelling change in some words that have final d(e). The addition of either ion or ive results in a change from de to s; e.g., divide + ion → division, explode + ive → explosive. "DE" is designated on a little more than 0.1% of the words in the lexicon.

The code "L" marks those words in which a final le (specifically, in the lexicon, the le of the suffixes acle, ible, and le) changes to il or ul with suffixation; e.g., (fab + le →) fable + ous → fabulous, (nob + le →) noble + ity → nobility, (poss + ible →) possible + ity → possibility, (spect + acle →) spectacle + ar → spectacular. "L" is designated on 0.2% of the words in the lexicon.

The code "R" marks a spelling change that occurs in some words with the er suffix. When another suffix is added, the vowel e in the er suffix is deleted from the spelling of the new word; e.g., anger + y → angry, enter + ance → entrance. (This process differs from that coded "RP" in which the vowel before r is deleted only in the pronunciation, not in the spelling.) "R" is designated on 0.3% of the words in the lexicon.

The code "S" marks words in which there have been spelling changes that are not covered by the other codes. Often the spelling change is one that occurs in such a few words that the specific change was not coded individually; e.g., abound + ance → abundance, generous + ity → generosity, permit + ion → permission. In some terms the change involves the addition of an epenthetic element; e.g., drama + ize → drama(t)ize, giant + ic → gi(g)antic, law + er → law(y)er. "S" is designated on approximately 2% of the words in the lexicon.

LINE 2 CODES

Line 2, Field 4: Correspondences, affixes, and epenthetic elements

Correspondences. The frequencies of the spelling-to-sound correspondences and the sound-to-spelling correspondences (listed and exemplified in Appendices C and E) are listed in Appendix I, Tables 1, 2, and 3. The frequencies of the exceptional spellings (listed and exemplified in Appendix D) are given at the end of each table in Appendix I.

The spelling-to-sound correspondence frequencies do not reflect frequencies of all grapheme units in the lexicon. The correspondences are identified only in bases, both free bases and bound bases. The letters that occur in prefixes, suffixes, and epenthetic elements are not identified apart from their units. Other codes have been employed to mark alterations in spelling and/or pronunciation of these morphemes.

The frequencies of consonant grapheme units, as shown in Appendix I, Table 1, range from over 3,000 for R and T to 6 for QUE and RH (see Table 15). Five consonants each have frequencies over 2,000; i.e., R, T, S, N, L. Although these figures represent occurrences in bases only, it is suspected that consonant frequencies in affixes are comparable. A cursory examination of the prefixes and suffixes in Appendices F and G indicates a relatively high frequency of occurrence of the consonants that rank in the top five on Table 15.

Some consonant letters would rank higher in a study of overall letter frequency. For instance, the assimilation process that occurs in the final consonant of many prefixes (e.g., ad + ford → afford, in + press → impress) increases the frequencies of the letters that result. Dewey's (1970) frequency study of English spellings does not discriminate affixes. Despite this difference between Dewey and the lexicon, a comparison of consonant frequencies in the two studies indicates many similarities. Although not in the same rank order, the top four consonants are the same, and the following six consonants are almost the same.

The frequencies of secondary vowels, as shown in Appendix I, Table 2, range from 443 for EA to 6 for EAU (see Table 16). There were 18 codes of "SV40" for sequences of two or more vowels, corresponding to a single sound, but not occurring frequently enough to be on the list of secondary vowels; e.g., EO as in leopard and people, IEU as in lieutenant.

The frequencies of primary vowels, as shown in Appendix I, Table 3, range from 3,200 for E to 154 for Y (see Table 17). Whereas most vowel codes represent sounds, the codes "E18", "E19", "E20", and "E30" identify final "silent" e's. These codes account for 43% of the total frequency for E. Deducting these nonsyllabic occurrences of final "silent" e, the remaining frequency for E is 1,814. Therefore, A, I, and O occur more often as "sounded" primary vowels than does E.

Table 15

Rank Order of Consonant Grapheme Units in Bases

Grapheme Units	Frequency Totals	Grapheme Units	Frequency Totals	Grapheme Units	Frequency Totals
R	3343	TH	323	FF	55
T	3218	CH	310	TCH	55
S	2411	LL	292	NN	53
N	2361	CK	276	Y	53
L	2033	SH	267	BB	37
P	1708	SS	189	DD	36
D	1696	J	176	DG	29
C	1653	RR	113	PP	29
M	1510	TT	101	SC	25
B	1141	NG	99	MM	23
G	901	X	95	GG	21
F	867	QU	94	ZZ	20
V	682	PH	87	CC	12
K	579	Z	86	GUE	9
H	567	WH	76	QUE	6
W	497	GH	63	RH	6

Table 16

Rank Order of Secondary Vowel Grapheme Units in Bases

Grapheme Units	Frequency Totals	Grapheme Units	Frequency Totals	Grapheme Units	Frequency Totals
EA	443	AU	83	OY	42
OU	340	IGH	78	UE	35
OO	278	EW	69	UI	30
EE	244	EY	66	OE	21
OW	213	IE	63	EU	9
AI	208	OI	61	YE	9
AY	122	AW	53	EAU	6
OA	117	EI	47		

Table 17

Rank Order of Primary Vowel Graphemes in Bases

Grapheme Units	Frequency Totals
E	3200
A	3106
I	2157
O	2022
U	1165
Y	154

Again, these frequencies do not represent all occurrences of the letters. All of the vowels are present in epenthetic elements and/or affixes. Also, all of the vowels can be found in sequences of two or more, as secondary vowels, in which they are considered only as components in a unit.

More analysis of all the data on spelling-to-sound correspondences is possible in order to explore further their relationships within the lexicon. The above comments only point out major features in terms of frequencies.

Affixes. The lexicon contains 1853 prefixes. Of these, 1820 are occurrences of the 34 identified prefixes (see Appendix I, Table 4); 73% are attached to bound bases (e.g., ex + tend → extend) and 27% are attached to free bases (e.g., ab + use → abuse). There were 33 codes of "PR" for 17 prefixes that did not occur frequently enough to be identified; e.g., the pur of purpose, the mega of megaphone.

The range of frequencies of prefixes is from 249 for CON to 5 for MONO and for MULTI (see Table 18). The three prefixes that rank highest in frequency are CON, RE, and AD--Latin prefixes that are almost always attached to Latin bases. The final consonants of both CON and AD are often assimilated (coded "9") with the first consonant of the base (see Appendix F). The prefix CON undergoes assimilation in 40% of its occurrences (e.g., con + pose → compose); AD is assimilated in 76% of its occurrences (e.g., ad + sist → assist). Together these two account for almost 60% of the prefix assimilation in the lexicon. Prefix assimilation takes place in 21% of the total occurrences of prefixes.

The lexicon includes 6158 suffixes. Of these, 6103 are occurrences of the 84 identified suffixes (see Appendix I, Table 5). There were 55 codes of "END" for 40 suffixes that did not occur frequently enough to be identified; e.g., the ace of terrace, the dom of freedom, the eer of engineer, the ique of unique, the ury of century.

Approximately half of these suffixes occur on bound bases (e.g., fort + une → fortune) and half occur on free bases (e.g., speak + er → speaker). Most suffixes occur after both bound bases and free bases. However, in the lexicon, the following 12 suffixes do not follow free bases: ACLE, EM, EST, IL, ILE, IN, OM, OP, OT, ULE, UNE, and UTE.

The range of frequencies of suffixes is from 945 for ER to 5 for each of six suffixes (see Table 19).

Epenthetic elements. There are 17 identified epenthetic element codes (see p. 25) with a total frequency of 361 (see Appendix I, Table 6). The code "EP" marks 5 different elements that occurred a total of 13 times, but not with enough frequency to be designated individually; e.g., et as in energetic, ig as in navigate, iv as in cultivate.

Table 18

Rank Order of Prefixes

Prefixes	Frequency Totals	Prefixes	Frequency Totals	Prefixes	Frequency Totals
CON	249	E	55	SUR	13
RE	200	PRE	44	FOR	12
AD	172	SUB	41	PARA	11
IN	134	BE	38	SUPER	10
DE	122	OB	38	SYN	10
EX	100	IN#	35	MIS	9
DIS	94	PER	22	TELE	8
A	86	INTER	21	DIA	7
UN	76	SE	19	AL	6
PRO	72	AB	17	ANTI	6
EN	68	TRANS	15	MONO	5
				MULTI	5

Table 19

Rank Order of Suffixes

Suffix	Frequency Totals	Suffix	Frequency Totals	Suffix	Frequency Totals
ER	945	IVE	47	ITE	19
Y	472	ANT	46	OM	19
ING	367	ABLE	41	ILE	18
ION	302	ARY	40	IST	17
LE	300	IA	39	EST	16
ED	287	ENCE	38	IBLE	16
LY	260	IE	38	TEEN	14
ATE	218	TH	36	SHIP	13
AL	204	ANCE	35	A	12
S	191	ERY	31	OT	12
EN	154	IS	31	RY	12
OR	146	IT	30	ADE	11
IC	128	NESS	28	ESS	11
ON	106	TY	27	IO	11
ENT	83	UM	27	EM	9
MENT	83	US	27	UNE	9
ET	82	LESS	25	ETTE	8
OUS	78	ORY	24	ISM	8
EL	74	ERN	23	LET	7
AN	65	ID	23	T	7
IN	62	IL	23	UDE	7
URE	59	WARD	23	OP	6
FUL	57	ICE	22	ACLE	5
ISH	54	IFY	22	EE	5
AGE	53	IZE	22	HOOD	5
ITY	53	ARD	21	SOME	5
INE	51	UTE	21	STER	5
AR	48	AIN	19	ULE	5

The range of frequency of epenthetic elements is from 62 for "I-" to 3 for "I1" followed by a comma (see Table 20). An epenthetic element that is followed by a hyphen connects a base and a suffix; one that is followed by a comma connects two components of a Latin or Greek compound. The three elements that occur most frequenctly are: "I-", which merges with and palatalizes the preceding consonant (e.g., patience, special), "I1-", in which the vowel is tense but unstressed (e.g., alien, stadium), and "AT1", in which the syllable is under primary or secondary stress (e.g., quotátion, spéctàtor).

Table 20

Rank Order of Epenthetic Elements

Code	Frequency	Code	Frequency	Code	Frequency
I-	62	IT1-	19	01,	10
I1-	55	I2-	17	U2-	9
AT1-	51	I,	14	I3-	7
IT-	31	AT-	11	IC1-	5
U-	24	IC-	11	I1,	3
0,	21	UL-	11		

Line 2, Field 5: Bases

Free bases. A free base is coded in this field only if it is part of a larger word. For example, the word suitable contains the free base suit; SUIT is coded in the base field. The two words cannot and can't each contain the free bases can and not; both CAN and NOT are coded in the the base field for cannot and for can't. However, the free base tree does not happen to occur in any other word in the lexicon, either with affixation or as a compound. Therefore, although tree is a listed word in the lexicon, it is not found in the field of bases. If all words having inflectional suffixes had been included, there would have been more bases listed, as well as more occurrences of some that are listed.

Items coded with free bases are of three kinds: compound words, affixed words, and contractions. There are 52 free bases that occurred with frequencies over 10 (see Table 21). The free bases used most

Table 21

Rank Order of Free Bases with Frequencies Over 10

Base	Frequency	Base	Frequency	Base	Frequency
MAN and MAN9	51	FIRE	17	STAND	13
OUT	39	LINE	17	BOOK	12
LAND and LAND9	32	TIME	17	FISH	12
OVER	32	BOARD	16	IS	12
UP	31	HAND	16	MAKE	12
WAY	26	DOWN	15	NORTH	12
DAY	25	GRAPH	15	WILL	12
MEN and MEN9	23	PART	15	CLOTH	11
HOUSE	22	SUN	15	CRACK	11
UNDER	22	TAKE	15	FORE	11
WOOD	21	COME	14	HIGH	11
NEW	19	FOOT	14	HOME	11
NOT	19	HEAD	14	MOUNT	11
AIR	18	JACK	14	SON	11
BACK	18	WORK	14	SOUTH	11
LIGHT	18	BALL	13	TOP	11
SELF	18	BOAT	13	USE	11
		SNOW	13		

frequently in compounds are MAN[20] (e.g., foreman), OUT (e.g., outline), LAND[20] (e.g., inland), OVER (e.g., overnight), and UP (e.g., upstairs). The frequency of LAND was almost doubled by the inclusion of geographic terms; e.g., Cleveland, England. Many more free bases occurred in compounds than in affixed forms. However, a few free bases had relatively high frequencies of affixation; PART occurred in 15 different affixed forms (e.g., depart, partial), and USE occurred in 11 affixed forms (e.g., abuse, useless). The frequencies of two free bases were higher than might be expected because they are contracted forms; NOT was contracted on 16 different verbal auxiliaries (e.g., isn't), and WILL was contracted on 6 personal pronouns (e.g., I'll).

Bound bases. The bound bases that are coded in this field are either Latin or Greek, and are only those that occur frequently in the vocabulary of childen (see Appendix H). Many bound bases do not occur frequently and are not identified on the list of bases. However, the code of "B" was assigned to significant, but unlisted, bases. There were almost 400 codes of "B" in this field.

Bound bases are, of course, usually affixed. Some bound bases occurred with rather high frequencies (see Table 22); VID occurred in 33 different affixed forms (e.g., evident, vision), POS occurred in 29 affixed forms (e.g., deposit, position), and TAIN occurred in 26 affixed forms (e.g., attain, continent). There are 44 bound bases that occurred with frequencies over 10.

Coded in the lexicon are approximately 1,500 different bases, including both free and bound forms. These 1,500 bases occur with a total frequency of about 4,500. Over 50% of the bases that are coded are free bases with frequencies of 1; e.g., DROWSE from drowsy, DRUG from drugstore.

Further analysis is needed to identify the character of the many base forms coded in the lexicon and to determine the processes of vocabulary growth and expansion that are represented.

---

[20]Both stressed ("MAN", "MEN", "LAND") and unstressed ("MAN9", "MEN9", "LAND9").

Table 22

Rank Order of Bound Bases with Frequencies Over 10

Base	Frequency	Base	Frequency	Base	Frequency
VID and VID9	33	GEN	16	CENT#	12
POS	29	MIN	16	DIC and DIC9	12
TAIN and TAIN9	26	CEED and CEED9	15	FORM#	12
SPECT and SPECT9	22	CEIV and CEIV9	15	FORT#	12
TEND# and TEND#9	21	CUR	15	MOT	12
DUC and DUC9	20	POL	15	PEND and PEND9	12
FER	18	FIN	14	SERV	12
MIT and MIT9	18	JECT	14	TRACT	12
PAR	18	LECT	14	CORD#	11
PORT#	18	SID	14	LAT	11
VERT and VERT9	18	STRUCT and STRUCT9	14	LEG#	11
FIC and FIC9	17	UN#	14	PRESS#	11
PLY	17	CID and CID9	13	STANT and STANT9	11
FECT	16	OR#	13	TERR	11
		RECT	13		
		VENT	13		

CONCLUSION

Over a period of several years, SWRL has worked on the preparation of this communication skills lexicon. The first, but far from simplest, task of compiling the lexicon involved consulting a variety of sources to prepare a set of over 10,000 words used by elementary school children. While not all of the words may be completely appropriate, and while some appropriate words may be missing, this vocabulary base is large enough for most nontechnical instructional and research uses at the elementary school level.

The second task, that of coding each word, was even more difficult and time-consuming. Every effort was made to produce a lexicon accurately coded with useful linguistic information. On the other hand, a completely error-free coding may not be possible because of the chance of human error in tabulating and describing thousands of words with hundreds of code choices. In addition, the codes themselves may be seen as inadequate or inaccurate in light of other or additional views of language. However, the codes provide an extensive and detailed indication of the complexity and potential of children's vocabulary.

The possible uses of this lexicon are multiple. It furnishes a powerful base that is relevant to elementary level instruction in all of the communication skills. But it is also a rich source of hypotheses and data for the researcher who is interested in investigating the nature and development of children's language. A few of the research applications are: (1) a study of the bases and affixes and the sequences in which they are learned, (2) a study of the nature of compounds and a comparison of compounds in this current lexicon with those found in lexicons compiled many years ago, (3) comparisons with the content of other lexicons, (4) a comparison of the grade levels for specific words in the lexicon with the measured difficulty of the words, and (5) studies involving bicultural comparisons.

The present report provides a rough surface analysis of the data generated by the lexicon and its codes. Further analysis can provide greater understanding of children's vocabulary. While the work of establishing this lexicon is complete, the even greater work of using it is only now beginning.

References

The American Heritage dictionary of the English language. New York: American Heritage Publishing Co., 1969.

Barnhart, C. L., Steinmetz, S., & Barnhart, R. K. (Eds.). The Barnhart dictionary of new English since 1963. Bronxville, New York: Barnhart/Harper & Row, 1973.

Berdiansky, B., Cronnell, B., & Koehler, J., Jr. Spelling-sound relations and primary form-class descriptions for speech-comprehension vocabularies of 6-9 year olds (Technical Report 15). Los Alamitos, California: SWRL Educational Research and Development, 1969.

Berdiansky, B., Stanton, G., & Cronnell, B. Design for sequencing spelling-to-sound correspondences for the SWRL Reading Program (Technical Report 47). Los Alamitos, California: SWRL Educational Research and Development, 1973.

Berg, P. C. A dictionary of new words in English. London: George Allen and Unwin, 1953.

Berger, K. W. An evaluation of the Thorndike and Lorge word count. Central States Speech Journal, 1971, 22, 61-64.

Berko, J. The child's learning of English morphology. Word, 1958, 14, 150-177.

Carroll, J. B., Davies, P., & Richman, B. The American Heritage word frequency book. New York: American Heritage Publishing Co., 1971.

Cronnell, B. Annotated spelling-to-sound correspondence rules (Technical Report 32). Los Alamitos, California: SWRL Educational Research and Development, 1971. (a)

Cronnell, B. Proper names for use in beginning reading (Technical Report 36). Los Alamitos, California: SWRL Educational Research and Development, 1971. (b)

Cronnell, B. Spelling-to-sound correspondences for reading vs. sound-to-spelling correspondences for writing (Professional Paper 12). Los Alamitos, California: SWRL Educational Research and Development, 1971. (c)

Dewey, G. D. Relative frequency of English spellings. New York: Teachers College Press, Columbia University, 1970.

Durr, W. K. A computer study of high-frequency words in popular trade juveniles. Paper presented at the International Reading Association, Anaheim, California, May 6-9, 1970. (ERIC ED 044 024)

Entwisle, D. R. Word associations of young children. Baltimore: The Johns Hopkins Press, 1966.

Francis, W. N. The structure of American English. New York: The Ronald Press Company, 1958.

Gates, A. I. A reading vocabulary for the primary grades. New York: Teachers College, Columbia University, 1935.

Green, G. I., Howard, J., Joerger, H. A., & Marino, M. A vocabulary list based on word frequency in fifth and sixth grade textbooks. Master's Thesis, Boston University, 1958.

Harris, A. J., & Jacobson, M. D. Basic elementary reading vocabularies. New York: Macmillan, 1972.

Horn, M, D. et al. A study of the vocabulary of children before entering the first grade. Baltimore: The Child Study Committee of the International Kindergarten Union, 1928.

Jacobs, H. D. A free-association word list for the Willamette Valley. Oregon School Study Council Bulletin, 1967, 11(3).

Kenyon, J. S., & Knott, T. A. A pronouncing dictionary of American English. Springfield, Mass: G. & C. Merriam Co., 1953.

Kolson, C. J. The vocabulary of kindergarten children. (Doctoral dissertation, University of Pittsburgh) Ann Arbor, Michigan: University Microfilms, 1960. No. 60-6184.

Kučera, H., & Francis, W. N. Computational analysis of present-day American English. Providence, Rhode Island: Brown University Press, 1967.

Murphy, H. A. and others. The spontaneous speaking vocabulary of children in primary grades. Journal of Education, 1957, 140, 1-105.

The New York Times encyclopedic almanac 1972. New York: The New York Times Company, 1971.

Reifer, M. Dictionary of new words. New York: Philosophical Library Inc., 1955.

Rinsland, H. D. A basic vocabulary of elementary school children. New York: Macmillan, 1945.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College, 1944.

Weaver, J. N. A study of word frequency in the oral vocabulary of kindergarten children. Master's Thesis, San Jose State College, 1955.

Webster's seventh new collegiate dictionary. Springfield, Mass: G. & C. Merriam Co., 1967.

Webster's third new international dictionary. Springfield, Mass: G. & C. Merriam Co., 1961.

Webster's third new international dictionary. Addenda section. Springfield, Mass: G. & C. Merriam Co., 1971.

Appendix A

NEW WORD STUDY

The communication skills lexicon is based on the most comprehensive and up-to-date studies of children's word use that are available. However, the major studies were published 20-30 years ago and include data gathered during many preceding years. The more recent studies are based on much smaller samplings and are primarily either classroom-oriented or are taken from reading materials. They do not reflect many lexical innovations in general use during the past two decades. To supplement this deficiency, a study was conducted to compile a list of words, relatively new in K-6 vocabulary, derived from a total of six valid compilations. Work was begun in March, 1972, and continued through April, 1974, as additional sources became available.

SOURCES

The first source used (Berg, 1953) represents an effort to collect all lexical items reflecting vocabulary growth and change: neologisms, acronyms, new phrases, new meanings added to old words, etc. The compiler recorded new terms from both oral and written language.

The second source consulted was also a glossary of new words and new word uses (Reifer, 1955). No information was available concerning either the sources for the material or the procedures followed in compiling the dictionary. However, it is a valuable and comprehensive study of relatively new terms.

The third source was a list of "New Words and Phrases" found in The New York Times Encyclopedic Almanac 1972 (pp. 428-430). These are also new terms plus new meanings for older terms, and were prepared by the lexicographical staff of The World Publishing Company.

The publication of the 1971 addenda section of Webster's Third New International Dictionary (1961) provided the fourth source. This list represents words that have, according to the publisher, G. & C. Merriam Company, become established in the language since 1961.

Berger (1971) conducted a study in which a sample of 10,000 words was taken from the November 10, 1969 edition of The New York Times. In a comparison with the Thorndike and Lorge (1944) word count, Berger found 78 words, relatively new in the language, that did not appear at all in the older study. This list of new words was the fifth source.

Barnhart, Steinmetz, and Barnhart (1973), with approximately 5000 entries, was the last source used. The new words and phrases were collected from United States, British, and Canadian writing--newspapers, magazines, and books. For this dictionary, terms were judged "new" if they had not appeared in any major standard English dictionary before 1963.

INCLUSION-EXCLUSION CRITERIA

Because of the variances in the sources, some additional inclusion-exclusion criteria were applied together with those defined for the communication skills lexicon (see Section I).

Words were included if they were judged to be known by K-6 American children, and if the meaning listed in the source was the only or primary meaning in use by children today; e.g., parka, penicillin.

Words were excluded if only the listed meaning was new, and there was another meaning, older but still common. For example, cartridge, referring to a case of film to be loaded into a camera, is relatively new; cartridge, referring to the case for a bullet, is an old meaning, but is still common. The word profile, meaning a short biographical sketch, is of fairly recent usage; profile meaning a side view of a human head, is not new, but is still common. Cartridge and profile were not included as new words. Also excluded were words which were new only in their uses as other parts of speech, not in their forms or meanings; e.g., headline as a verb.

Some words were not included because of their transitory nature or because they are educationally unacceptable. These exclusions consisted of: (1) words--primarily slang--that seem to be changing now (e.g., kook, mooch), (2) words that refer only to fashions and are created at the capricious whim of designers and advertisers (e.g., mini, midi, maxi, pantdress), and (3) words--especially from the drug culture--that might be offensive or objectionable (e.g., acidhead, bad, bennies, narc).

PROCEDURES

Each source was reviewed by two staff members, who listed only those words in current use by children. The two lists were compared for agreement; those words lacking agreement were reconsidered for inclusion. The two resulting lists from each source were combined.

This list was edited independently by eight SWRL staff members, selected for their knowledge of children's contemporary vocabulary. They were asked to cross out words not currently known by K-6 children. The term remaining were those that had been agreed upon by at least 50% of the eight staff members.

After all of the available sources had been reviewed and edited, 7% of the listed terms seemed to be reliably well-known and are included in the lexicon; e.g., ad, cheeseburger, drive-in, sweatshirt.

An additional validation procedure was employed in which the remaining 93% of the terms were divided into five lists of approximately 80 words each. All five lists were sent to each of seven sixth-grade public school classes having students of average ability and the lists

were distributed equally within each class. There were 166 students who participated. Teachers administered the word lists and were instructed to provide help only in reading the words, not in interpreting the meanings. Children were asked to (1) circle a word that was known and (2) draw a line through a word that was not known. The children were allowed only 30 minutes in which to edit the list.

Words circled by 60% of the responding students are included in the lexicon of words known by children in the K-6 range. This validation procedure netted 293 new words. These, added to the 29 reliably well-known words that were not tested, constitute a final list of 322 new words.

Appendix B

SYMBOLS

Pronunciation symbols

Vowels

Symbol	Examples
/i/	neat, see
/ɪ/	bit, milk
/e/	name, day
/ɛ/	get, head
/æ/	fat, bad
/a/	hot, car
/ɔ/	ball, loss
/o/	bone, go
/u/	push, book
/u/	food, tune
/ʌ/	but, love
/ə/[1]	around, elephant
/ay/	cry, mine
/aw/	found, owl
/oy/	boy, noise

Consonants

Symbol	Examples
/b/	boy, cab
/č/	chip, hatch
/d/	dead, do
/f/	fun, off
/g/	go, egg
/h/	home, head
/ǰ/	judge, age
/k/	kick, cat
/l/	let, hill
/m/	man, ham
/n/	no, hand
/ŋ/	sing, think
/p/	pull, trip
/r/	red, far
/s/	see, miss
/š/	she, hash
/t/	ten, hit
/θ/	thing, breath
/ð/	them, breathe
/v/	vase, love
/w/	wet, quick
/y/	yet, you
/z/	zoo, please
/ž/	azure, treasure

[1] The sound /ə/ is found only in unstressed syllables. It often alternates with /ɪ/, but only /ə/ is used in this report.

Notational Devices

/ /	pronunciation
→	"is pronounced as"
C	any consonant
C_{vd}	any voiced consonant (e.g., /b, d, m, l, v/)
V	any vowel
$\bar{V}$	a tense vowel sound (/e, i, ay, o, (y)u/)
$\breve{V}$	a lax vowel sound (/æ, ɛ, ɪ, ə, ʌ/)
∅	no sound
/__	"in the environment of"
( )	optional
{ }	alternative
#	base word boundary

Appendix C

CODES FOR SPELLING-TO-SOUND CORRESPONDENCES

Grapheme Unit	Coding Number	Description (Environments in parentheses are defining, not predicting.)	Examples (grapheme unit underlined)
Consonants			
B		b → /b/	boy, crib, number
	5	b → Ø /m__#/	climb, comb
BB		bb → /b/	bubble, hobby
C	1	c → /s/ /__ {e, i, y}	lace, city, fancy
	2	c → /k/ /__ {a, o, u, C, #}	cat, come, cup, cream, disc
	8	c → /š/	social, ancient
CC		cc → /k/	soccer, raccoon
CH		ch → /č/	cheap, church
	6	ch → /k/	ache, school, chord, chrome
	7	ch → /š/	chef, machine
CK		ck → /k/	kick, packet
D		d → /d/	dead, needle
	9	d → /j/	educate, cordial
DD		dd → /d/	sudden, paddle

DG		dg → /ǰ/	judge, ledge
F		f → /f/	fat, after
FF		ff → /f/	off, taffy
G	1	g → /ǰ/ / _ {e, i, y}	gem, age, giant gypsy
	2	g → /g/ / _ {a, o, u, C, #}	gave, go, gum, green, bag
	5	g → Ø / {#_n, _n#}	gnaw, sign
	6	g → /g/ (/ _ {e, i, y})	get, girl
	9	g → /ǰ/ (before a suffix)	religion, region
GG		gg → /g/	egg, wiggle
GH		gh → Ø	through, taught
GUE		gue → /g/ / _ #	league, vague
H		h → /h/	home, ahead
	5	h → Ø	hour, graham, bah
J		j → /ǰ/	joy, judge
K		k → /k/	milk, kill
	5	k → Ø / #_n	know, knot
L		l → /l/	like, mile
	5	l → Ø / {o_k, a_f, a_k, a_m, a_v}	yolk, half, walk, palm, calves

LL		ll → /l/	fill, bullet
M		m → /m/	man, came
MM		mm → /m/	summer, mammal
N		n → /n/	no, nine
	4	n → /ŋ/ /__ {/g/ /k/}	thank, anxious, uncle, banquet, finger
	5	n → Ø /m_#	hymn, autumn
	6	n → /n̩/ (syllabic)	isn't
NG		ng → /ŋ/ /__#	ring, song
NN		nn → /n/	manner, penny
P		p → /p/	people, pop
PH		ph → /f/	phone, photo
PP		pp → /p/	pepper, apple
QU		qu → /kw/	quick, banquet
	6	qu → /k/ (not /__e#)	conquer, bouquet
QUE		que → /k/ /__#	mosque, antique
R		r → /r/	run, far
RH		rh → /r/	rhyme, rhubarb
RR		rr → /r/	hurry, marry
S		s → /s/ (not /V__V)	seen, fast, horse
	2	s → /s/ / {a ea oo ou} __e#	case, lease, moose, house
	3	s → /z/ /V__V	nose, cheese, music
	4	s → /z/ / {__Cvd Cvd__}	measles, pansy
	6	s → /z/ /V__V (exceptions to S3)	dose, geese, basin

S (cont.)	7	s → /z/ /__# (not a suffix)	is, has
	8	s → /ž/ (before a suffix)	pension, mansion
	9	s → /ž/	usual, measure
SC		sc → /s/ /__{e, i, y}	scene, scissors
SH		sh → /š/	shoe, rush
SS		ss → /s/	miss, message
	8	ss → /š/	issue, mission
T		t → /t/	title, let
	5	t → ∅ /{s__le, s__en, f__en}	wrestle, fasten, often
	8	t → /š/	nation, partial
	9	t → /č/	statue, nature
TCH		tch → /č/	match, butcher
TH	1	th → /θ/	thin, bath
	2	th → /ð/ /V__y	bathe, father
	3	th → /ð/ only in function words	they, although, the
TT		tt → /t/	mitt, little
V		v → /v/	vase, love
W		w → /w/	wet, beware
	5	w → ∅ /#__r	wren, wrong
WH		wh → /hw/, /w/	when, whether

X		x → /ks/	box, axis
	8	x → /kš/	anxious, complexion
Y	1	y → /y/	yet, beyond
Z		z → /z/	zoo, lazy
ZZ		zz → /z/	buzz, fuzzy

Secondary vowels

AI		ai → /e/	stain, air
AU		au → /ɔ/, /a/	cause, Laura
AW		aw → /ɔ/, /a/	saw, hawk
AY		ay → /e/	day, play
EA		ea → /i/	each, hear
	6	ea → /ɛ/	bread, bear
	7	ea → /ʌ/ (/__r)	earn, search
EAU		eau → /o/	bureau, plateau
EE		ee → /i/	beet, deer
EI	1	ei → /e/ /__{gn, n, gh, r}	reign, rein, neighbor, heir
	2	ei → /i/	ceiling, receive
EU		eu → /(y)u/	neutral, feud, Europe
EW		ew → /(y)u/	few, new
EY		ey → /e/	they, obey
	5	ey → /i/ in unstressed syllables	donkey, monkey
IE	1	ie → /ay/ /__#	die, lie
	2	ie → /i/ /__ ...	field, fierce
IGH		igh → /ay/	high, right
OA		oa → /o/	load, roar
OE		oe → /o/ /__#	toe, hoe
OI		oi → /oy/	noise, join

OO	1	oo → /u/	br<u>oo</u>m, p<u>oo</u>r
	2	oo → /u/	b<u>oo</u>k, w<u>oo</u>d
OU		ou → /aw/	c<u>ou</u>nt, s<u>ou</u>r
	6	ou → /ʌ/	c<u>ou</u>ple, j<u>ou</u>rney
	7	ou → /(y)u/	thr<u>ou</u>gh, gr<u>ou</u>p, t<u>ou</u>r
	8	ou → /ɔ/, /a/ (/__gh)	f<u>ou</u>ght, th<u>ou</u>ght
	9	ou → /o/	s<u>ou</u>l, th<u>ou</u>gh, p<u>our</u>
OW	1	ow → /o/	gl<u>ow</u>, bel<u>ow</u>
	2	ow → /aw/	n<u>ow</u>, all<u>ow</u>
OY		oy → /oy/	b<u>oy</u>, t<u>oy</u>
UE		ue → /(y)u/	bl<u>ue</u>, arg<u>ue</u>
UI		ui → /(y)u/	fr<u>ui</u>t, j<u>ui</u>ce
YE		ye → /ay/	l<u>ye</u>, b<u>ye</u>

Primary vowels

A	11	a → /e/ /__Ce#	name, care
	12	a → /e/ /__C $\{^{r}_{l}\}$e	acre, stable
	13	a → /e/ /__CV	baby, vary
	14	a → /e/ /__C $\{^{r}_{l}\}$ V	April, fragrant
	15	a → /æ/ /__C(C)(C)#	sat, fast
	16	a → /æ/ /__CC...	saddle, jacket
	17	a → /ə/, /ɪ/ in unstressed syllables (/_....;) not A25)	compass, career
	18	a → /e/ /__ $\{^{nge}_{ste}\}$	strange, paste
	20	a → /e/, /ɛ/, /æ/ /_rr	marry, carry
	21	a → /ɑ/ /__r $\{^{C}_{\#}\}$	cart, car
	22	a → /o/ //w/__r	wharf, war, quart
	23	a → /ɔ/, /ɑ/ /__ $\begin{Bmatrix} ll\# \\ lk \\ lt \\ ld \end{Bmatrix}$	ball, walk, salt, bald
	24	a → /ɑ/, /ɔ/ // /__	wad, squat
	25	a → /ə/ /__#	comma, sofa
	26	a → /e/ /__ V	cacao
	27	a → /æ/ /__C $\begin{Bmatrix} ic \\ id \\ ish \\ it \\ ity \end{Bmatrix}$ #	magic, rapid, radish, habit, capacity
	28	a → /æ/ /__x V	axis, taxi
	29	a → /æ/ /'__ CV(C)(C)(C)V...	animal, majesty
	32	a → #	separate (adj), diamond

A (cont.)	33	a → /æ/ /__C($\begin{Bmatrix} r \\ l \end{Bmatrix}$)V	cabin, talent
	34	a → /ɑ/ (not /$\begin{Bmatrix} __l \\ __r \\ w__ \end{Bmatrix}$)	drama, father
	35	a → /ɔ/, /ɑ/ /__l... (not A23)	walnut, false
	39	a → /e/ /__$\begin{Bmatrix} ch \\ ph \\ th \end{Bmatrix}$V	ache, aphid, bathe
E	11	e → /i/ /__Ce#	scene, here
	13	e → /i/ /__CV	meter, hero
	14	e → /i/ /__C$\begin{Bmatrix} r \\ l \end{Bmatrix}$V	zebra, cathedral
	15	e → /ɛ/ /__C(C)(C)#	set, felt, tent
	16	e → /ɛ/ /__CC...	edge, merry
	17	e → /ə/, /ɪ/ in unstressed syllables	elephant, solemn
	18	e → Ø /VC(C)__# (when V → /V̄/ and C ≠ c,g,s,th)	name, tone
	19	e → Ø /V(C)C__# (when V → /V̄/ and C = c,g,s,th)	face, age, nose, bathe
	20	e → Ø /C__# (when not VC(C)e → /V̄C(C)/ and C = c,g,s,th,v,z)	choice, edge, immense, breathe, serve, breeze
	21	e → /ʌ/ /__r$\begin{Bmatrix} C \\ \# \end{Bmatrix}$	her, herd
	25	e → /i/ /#/C/__#	she, me
	26	e → /i/ /__V	create, theater
	27	e → /ɛ/ /__C$\begin{Bmatrix} ic \\ id \\ ish \\ it \\ ity \end{Bmatrix}$#	relic, cherish, credit

E (cont.)	28	e → /ɛ/ /__xV	Texas, Mexico
	29	e → /ɛ/ /'__CV(C)(C)(C)V...	genuine, element
	30	e → Ø /C__# (not E18, E19, E20)	giraffe, Maine
	32	e → Ø	difference, several
	33	e → /ɛ/ /__C({r, l})V	metal, senate
	35	e → /i/ /C__#	recipe, adobe
	39	e → /i/ /__{ch, ph, th}V	Stephen
I	11	i → /ay/ /__Ce#	like, fire
	12	i → /ay/ /__C{r, l}e#	title, idle
	13	i → /ay/ /__CV	pilot, cider
	14	i → /ay/ /__C{r, l}V	migrate, vibrate
	15	i → /ɪ/ /__C(C)(C)#	sit, milk
	16	i → /ɪ/ /__CC...	little, village, mirror
	17	i → /ə/, /ɪ/ in unstressed syllables	evidence, admiral
	18	i → /ay/ /__{nd, ld, gn}#	find, wild, sign
	21	i → /ʌ/ /__r{C, #}	bird, sir
	25	i → /i/ /__#	ski, Miami
	26	i → /ay/ /__V	diet, giant
	27	i → /ɪ/ /__C{ic, id, ish, it, ity}#	clinic, timid, finish, visit, utility
	29	i → /ɪ/ /'__CV(C)(C)(C)V...	citizen, minister
	32	i → Ø	aspirin, family

I (cont.)	33	i → /ɪ/ /__C({r, l})V	city, lizard
	35	i → /i/ /__C(u)e# (not INE)	antique, suite
	36	i → /i/ /__V	period, piano
	37	i → /i/ (not I25, I35, I36)	pizza, tortilla
	39	i → /ay/ /__ {ch, ph, th} V	Michael
O	11	o → /o/ /__Ce#	home, snore
	12	o → /o/ /__C{r, l}e#	ogre, noble
	13	o → /o/ /__CV	odor, glory
	14	o → /o/ /__C{r, l}V	okra, only
	15	o → /a/ /__C(C)(C)#	lot, lock
	16	o → /a/ /__CC...	motto, hockey
	17	o → /ə/, /ɪ/ in unstressed syllables	method, corral
	21	o → /o/, /a/ /__r{C, #}	horn, tortoise
	22	o → /ʌ/ /w__rC	word, worth
	23	o → /ɔ/ /__l(C)#	roll, gold, yolk, boil, control
	24	o → /ɔ/, /a/ /__{fC, ng, ss, st, th, g}#	off, soft, long, loss lost, moth, dog
	25	o → /o/ /__#	go, hobo
	26	o → /o/ /__V	poet, oasis
	27	o → /a/ /__C{ic, id, ish, it, ity}#	comic, solid, astonish, deposit, velocity
	28	o → /a/ /__xV	antitoxin, oxygen

O (cont.)	29	o → /a/ /'__CV(C)(C)(C)V...	comedy, monument
	31	o → /ʌ/	love, mother
	32	o → ∅	chocolate
	33	o → /a/ /__C({r 1})V	body, honest
	34	o → /o/ /__st	post, most
	39	o → /o/ /__ {ch ph th} V	clothe, trophy
U	11	u → /(y)u/ /__Ce#	use, dune, cure
	12	u → /(y)u/ /__C{r 1}V#	bugle
	13	u → /(y)u/ /__CV	unit, crusade, fury
	14	u → /(y)u/ /__C{r 1}V	nuclear, lubricate
	15	u → /ʌ/ /__C(C)(C)#	gum, fuss
	16	u → /ʌ/ /__CC...	sudden
	17	u → /(y)ə/, /(y)ɪ/ in unstressed syllables	stirrup, porcupine
	20	u → ∅ /#g__V	guest, guard
	21	u → /ʌ/ /__r{C ∅}	fur, burn
	24	u → /w/	language, jaguar
	26	u → /(y)u/ /__V	fluid, ruin
	28	u → /ʌ/ /__xV	luxury
	31	u → /ʊ/	put, push
	33	u → /ʌ/ /__C({r 1})V	study, public
Y	11	y → /ay/ /__Ce#	rhyme, type
	12	y → /ay/ /__C{r 1}e#	cycle
	13	y → /ay/ /__CV	nylon, tyrant
	14	y → /ay/ /__C{r 1}V	cyclone, hydrant

Y (cont.)	15	y → /ɪ/ /__C(C)(C)#	myth, gym
	16	y → /ɪ/ /__CC...	system, gypsy
	17	y → /ə/, /ɪ/ in unstressed syllables	oxygen, mysterious
	25	y → /ay/ /'__#	deny, try
	26	y → /ay/ /__V	hyacinth, Wyoming
	27	y → /ɪ/ /__C $\left\{\begin{matrix}\text{ic}\\\text{id}\\\text{ish}\\\text{it}\\\text{ity}\end{matrix}\right\}$ #	physics
	29	y → /ɪ/ /__CV(C)(C)(C)V...	sycamore, cylinder
	33	y → /ɪ/ /__C($\{\begin{matrix}\text{r}\\\text{l}\end{matrix}\}$)V	bicycle, syrup
	36	y → /i/ /__V	Tokyo, polyester
	39	y → /ay/ /__ $\left\{\begin{matrix}\text{ch}\\\text{ph}\\\text{th}\end{matrix}\right\}$ V	psychology, typhoon, scythe

Appendix D

CODES FOR EXCEPTIONAL SPELLING-TO-SOUND CORRESPONDENCES

			Example
40:	miscellaneous, infrequent pronunciations, one or two words only		any, waltz
45:	Ø (not included in other correspondence)		debt, circuit

Consonants

G	41	/ž/	rouge
GH	41	/g/	ghost
	42	/f/	laugh
NG	41	/ŋ/ (not /__#)	tongue
S	42	/z/ (exceptions to S2)	phase, please, choose, arouse
SS	41	/z/	scissors
TH	41	/θ/ V__V (exceptions to TH2)	method
	42	/ð/ (not TH2, TH3)	smooth
	43	/t/	Thomas
WH	42	/h/	who

Secondary Vowels

AU	41	/æ/	laugh
EA	41	/e/	break
	42	/a/ /__r	heart
EY	41	eye → /ay/ (used as code for whole word)	
OA	41	/ɔ/, /a/	broad

OO	41	/o/ /__r	door
	42	/ʌ/	flood
UI	42	/ɪ/ (stressed)	build

SV40 -- Any sequence of two or more vowels, corresponding to a single sound, but not on secondary vowel list (e.g., eo, aa, uu, ieu)

Vowels

41: V → /V̆/ /_Ce# have, Adele

46: V → /V̄/ /_CC chamber, pint

47: V → /ɪ/, /ə/ /_Ce# furnace, lettuce

49: V → /V̆/ /_ {ph, th, ch} V bother, nephew

A	48	/æ/ /_ {ngɛ, ste} (not A18)	anger, caste
E	42	/e/	crepe
	43	/e/ /__r (e)	where
	44	/ɪ/	pretty
I	42	/ɪ/ /_ {nd, ld}	wind (noun), wilderness
O	42	/u/ (not /__v/)	to
	43	/u/ /__v	move
	44	/ʊ/	woman
	48	/ɔ/, /a/ (not 024)	gone

Appendix E

ADDITIONAL CODES FOR SPELLING-TO-SOUND CORRESPONDENCES

Sound-to-spelling correspondences

all doubled consonants (+):	/__#, /#__	egg, bell, llama
C1+ :	not /__e#	cent, city
C1# :	/C__e#	dance
C1@ :	/VV__e#	piece
C2+ :	/__#, /__le#	disc, uncle
CH+ :	/CV__#	rich, sandwich
CK+ :	not //$\acute{\breve{V}}$/__	Bismarck, hammock
DG+ :	not //$\acute{\breve{V}}$/__	porridge
F+ :	/CV__#	if
G1+ :	/CV__e# (V ≠ /V/), /_ {i, y, e...}	college, giraffe, gym, gem
G1# :	/C__e#	large
G1@ :	/VV__e#	siege
K+ :	/__ {a, o, u, C} , /CV__# (C ≠ le#)	kangaroo, Korea, skunk, okra, trek
L+ :	/$\acute{CV}$__#	pal, control
S+ :	/$\acute{CV}$__#	bus
S# :	/C__e#	horse
S3@ :	/VV__e#	cheese
S6@ :	/VV__e#	geese
SS@ :	not /$\acute{V}$_#	compass

Note: Symbols may be concatenated, e.g., "#+".

V+	: /__#	Bev
V#	: /C__e#	valve
V@	: /VV__e#	leave
Z+	: /__#	quiz
Z#	: /C__e#	bronze
Z@	: /VV__e#	seize
OA@, OO1@, QU7@	: /__#	cocoa, moo, caribou
AW(#)@	: not /__ {#, V..., k#, l#, n#}	Lawrence
AY@	: not /__ {#, V...}	Wayne
EW@	: not /__ {#, V..., n#}	pewter
OU@	: /_ {l#, n#}	foul, noun
OWl@	: not /__#	bowl
OW2@	: not /__ {#, V..., l#, n#}	crowd, growl
OY@	: not /__ {#, V...}	oyster

Phonological features

all vowels (#)	: /__r (unless the r is listed in the correspondence description) (When "#" is used, the vowel pronunciation may differ somewhat from that indicated in Appendix C.)	hair, more
all vowels with optional /y/ (+)	: /yu/	cute, few

Graphemic features

A33+, E33+, I33+, O33+, U33+, Y33+	: /__C $\left\{\begin{matrix}r\\l\end{matrix}\right\}$ V	Africa, metropolitan, citrus, geography, publish, bicycle
all single consonants (%): initial in a bound base when preceded by a prefix which ends with the same letter.		ommit, approve
all double consonants (%):	/ V (unstressed) ___	shellac
	/__e#	giraffe
	/__ $\left\{\begin{matrix}ic\\id\\ish\\it\\ity\end{matrix}\right\}$	rubbish

Appendix F

CODES FOR PREFIXES

Code	Pronunciation	Example
A	/ə/ (unstressed)	afoot
AB	/ǽb/	absent
AB2	/əb/	abuse
AD	/ǽd/	advertise
AD9	/ǽC/	accent, exaggerate, ally, appetite
AD2	/əd/	adore
AD92	/əC/	accuse, affect, aggression, allow, announce, appear, assign, attack
AD3	/ə/ (d merges with j)	adjust
AL#	/ɔl/ or /al/	already
ANTI	/ǽnti/	antifreeze
BE	/bɪ/	become
CON	/kǽn/	convent
CON9	/kǒ/ or /kǽC/	cohost, college, common, correspond
CON91	/kʌ́m/	comfort
CON2	/kən/	continue
CON92	/ko/ or /kəC/	cooperate, collect, compare, correct
CON3	/kǽŋ/	congress
DE	/dɪ́/	decomposition
DE1	/dɛ́/	definite
DE2	/dɪ/	define
DIA	/dáyə/	dialogue
DIA1	/dayǽ/	diameter
DIS	/dɪ́s/	discount
DIS9	/dɪ́/ or /dɪ́C/ or /dáy/	distance, differ, digest
DIS2	/dɪs/	discuss
DIS92	/də/	divide
DIS3	/dɪz/	dissolve
E	/ɛ́/	educate
E2	/ɪ/	erase

Note: The degree of stress on a prefix may vary in different words. and may change with further affixation.
Prefixes shown without stress marks may have vowel variations.
C represents an assimilated consonant and also indicates a spelling change.

Code	Pronunciation	Example
EN	/ɛ́n/	engine
EN9	/ɛ́C/	empire
EN2	/ɪn/	entire
EN92	/ɪC/	employ
EX	/ɛ́ks/	expert
EX9	/ɛ́C/	effort, escort (noun)
EX2	/ɪks/	expect
EX92	/ɪC/	effect, escape
EX3	/ɪgz/ or /ɛ́gz/	examine, exile
FOR	/fər/	forgive
IN	/ín/ ("inside")	inland
IN9	/íC/	illustrate, impulse, irritate
IN2	/ɪn/ ("inside")	inspect
IN92	/ɪC/	illuminate, impress
IN#	/ín/ ("not")	innocent
IN#9	/íC/	ignorant
IN#2	/ɪn/ ("not")	insane
IN#92	/ɪC/	ignore, impossible, irregular
INTER	/ìntər/	interfere
INTER9	/ìntəC/	intellectual
INTER1	/ɪntə́r/	interpret
INTER91	/ɪntɛ́C/	intelligent
INTER2	/íntr/	interest
MIS	/mɪs/	mischief
MONO	/mánə/	monoplane
MONO1	/mənɑ́/	monopoly
MULTI	/mʌ́ltə/	multiply
OB	/ɑ́b/	oblong
OB9	/ɑ́C/ or /ɔ́C/	occupy, offer, opposite
OB2	/əb/	object (verb)
OB92	/o/ or /əC/	omit, occasion, offend, oppose
PARA	/pærə/	paragraph
PARA1	/pərǽ/	paralysis
PER	/pʌ́r/	permanent
PER2	/pər/	permission

Code	Pronunciation	Example
PRE	/prɪ́/	premium
PRE1	/prɛ́/	president
PRE2	/prɪ/	predict
PRO	/prṓ/	program
PRO1	/prɑ́/	process
PRO2	/prə/	proceed
RE	/rɪ́/	recent
RE1	/rɛ́/	refuge
RE2	/rɪ/	review
SE	/sɪ́/	secret
SE1	/sɛ́/	secretary
SE2	/sɪ/	secure
SUB	/sʌ́b/ or /səb/	submit, suburb
SUB9	/sʌ́C/ or /səC/	succeed, suffer, suggest, support, suspect
SUPER	/sú̇per/	supervision
SUR	/sʌ́r/ or /sər/	surface, surround
SYN	/sɪ́n/ or /sɪn/	synonym, synthetic
SYN9	/sɪ́C/	syllable, sympathy
TELE	/tɛ́lə/	telephone
TRANS	/tráns/ or /træns/	transform, transit
UN	/ʌ́n/ or /ən/	untie

Appendix G

CODES FOR SUFFIXES

Code	Pronunciation	Example
A#	/ə/	harmonica
ABLE	/əbəl/	suitable
ACLE	/əkəl/	spectacle
ADE	/ɛ́d/	blockade
AGE	/ɪj̆/	shortage
AIN	/ən/	mountain
AL	/əl/	magical
AN	/ən/	Roman
ANCE	/əns/	assistance
ANT	/ənt/	servant
AR	/ər/	polar
ARD	/ərd/	custard
ARY	/rɪ/	boundary
ARY2	/əri/	rotary
ARY3	/ɛ̀ri/	missionary
ATE	/ə̀t/	decorate
ATE2	/ət/	fortunate
ED	/t/	fixed
ED2	/d/	alarmed
ED3	/əd/	crowded
ED4	/əd/ (unpredictable)	naked
EE*	/ī/	committee
EL	/əl/	novel
EM	/əm/	anthem
EN#	/ən/	blacken
ENCE	/əns/	existence

Note: The degree of stress on a suffix may vary in different words, and may change with further affixation. /ɪ/ and /ə/ alternate in many suffixes.

*Code used only for suffix added to free base.

Code	Pronunciation	Example
ENT	/ənt/	different
ER	/ər/	winter
ERN	/ərn/	cavern
ERY	/ri/ or /əri/	creamery
ESS	/əs/	duchess
EST	/əst/	honest
ET	/ət/	market
ET2	/e̊/	buffet
ETTE	/ɛ̊t/	cigarette
FUL	/fəl/	helpful
HOOD	/hu̇d/	childhood
IA	palatization + /ə/	militia
IA1	/yə/	ammonia
IA2	/iə/	India
IBLE	/əbə!/	possible
IC	/ɪk/	atomic
ICE	/əs/	service
ID	/əd/	acid
IE	/i/	auntie
IFY	/əfa̋y/	purify
IL	/əl/	peril
ILE	/əl/	hostile
iN@	/ən/	cabin
INE	/a̋yn/	valentine
INE2	/ən/	engine
INE3	/i̊n/	marine
ING	/ɪŋ/	building
IO	/io/	radio
ION	palatalization + /ən/	digestion
ION1	/yən/	onion
ION2	/iən/	accordion
IS	/əs/	crisis
ISH	/ɪš/	reddish
ISM	/ɪ̀zəm/	communism
IST	/əst/	artist

Code	Pronunciation	Example
IT@	/ət/	credit
ITE	/ə́yt/	polite
ITE2	/ət/	granite
ITY	/ətɪ/	purity
IVE	/əv/	native
IZE	/ə̀yz/	civilize
LE	/əl/	handle
LESS	/ləs/	helpless
LET	/lət/	booklet
LY	/lɪ/	mainly
MENT	/mənt/	assignment
NESS	/nəs/	kindness
OM	/əm/	bottom
ON	/ən/	dragon
ON2	/ə̀n/	rayon
OP	/əp/	gallop
OR	/ər/	actor
ORY	/ò̞rɪ/	observatory
ORY2	/rɪ/ or /ərɪ/	ivory
OT	/ət/	patriot
OUS	/əs/	generous
RY	/rɪ/	poetry
S11	/s/	politics
S12	/z/ (spelled s)	besides
S13	/əz/	riches
S14	/z/ (spelled es)	goes
SHIP	/x̀̀ɪp/	sportsmanship
SOME	/səm/	lonesome
STER	/stər/	youngster
T*	/t/	weight
TEEN	/tɪ̂n/	sixteen

*Code used only for suffix added to free base.

Code	Pronunciation	Example
TH*	/θ/	sixth, length
TY	/ti/	sixty
UDE	/ùd/	longitude
ULE	/(y)əl/ or palatalization + /ʊl/	capsule, schedule
UM	/əm/	album
UNE	/(y)ùn/ or palatalization + /ən/	Neptune, misfortune
URE	palatalization + /ər/	moisture
URE2	/yər/	figure
US	/əs/	minus
UTE	/ůt/	salute
UTE2	/yʊt/	distribute
WARD	/wərd/	backward
Y	/i/	hilly

*Code used only for suffix added to free base.

Appendix H

BOUND BASES

ANIM	animal
ARCH	architect
AST(E)R	astronaut
AUD	audience
AUTO	autograph
cap (CAPT9)	capital
CAPT	captain
CED	precede
CEED	proceed
CEIV	receive
CENT#	percent
ce(i)pt (CEIV9)	reception, receipt
CERT	concert
cess (CED9)	recess
cess (CEED9)	procession
CID	decide
CIP	principal
CIRC	circular
cis (CID9)	decision
CIT	recite
CIV	civil
CLIN	decline
CLUD	conclude
clus (CLUD9)	conclusion
COM	comic
CORD#	accord
CORP	corporation
COVER#	discover
CULT	culture
CUR	secure
CUS	accuse
CYC	cycle
DEM	democracy
DIC	indicate
DICT	predict
dit (DIC9)	condition
DOM	dominate
DUC	produce
duct (DUC9)	production
ELECTR	electric
EQU	equal
FACT#	factory
FECT	perfect
FEND	offend
fens (FEND9)	offense
FER	prefer
FESS	profess
FIC	efficient
FIN	define
fit (FIC9)	benefit
FORM#	inform
FORT#	comfort
FUS	confuse
GEN	general
GEST	digest
GN	recognize
GRAM	diagram
GRAPH	photograph
GRAT	gratitude
GRESS	progress
HAP	happen
HON	honest
HORR	horror
HUM#	human
HYDR	hydrogen
IDENT	identify
IM	image
IT#	exit
JECT	project
LAT	translate
LECT	elect
LEG#	legend
LEV	elevate
li (LY)	reliable
LIBER	liberty
LIEF	relief
liev (LIEF9)	relieve
LIG	religion
LOC	locate
LOG# (logue)	apology, dialogue
LY	ally
MAGN	magnet
MAJ	major
MAN#	manicure
MEDI (medy)	medium, remedy
MEM	memory
MERC	mercury
merch (MERC9)	merchant
MET	diameter
MICRO	microphone
MIN	minus
MIR	admire
miss (MIT9)	permission
MIT	permit
MOD	model
MONST(E)R	demonstrate
MOT	promote
MUN	community
MUS	music
NAT	natural
NAV	naval
NOM	nominate

NOUNC	announce
NUM	numerous
ONYM	synonym
OPER	operate
OR#	oral
ORD	order
PAR	separate
peat (PET9)	repeat
PED	centipede
PEL	propel
PEND	depend
pens (PEND9)	expensive
PERI	experiment
PET	petition
PHON	symphony
PLEX	complex
pli (PLY)	compliment
PLOR	explore
PLY	reply
POINT#	appoint
POL	police
PORT#	export
POS	expose
POSS	possible
PRESS#	express
PRIN	principal
PRIV	private
PUBL	public
PUT#	dispute
QUIR	inquire
quis (QUIR9)	exquisite
RAD	radio
RECT	correct
REG	region
RUPT	erupt
S	present
SCEND	ascend
scent (SCEND9)	descent
SCI	science
SCRIB	describe
script (SCRIB9)	description
SEMB	resemble
SENT	sentiment
SERT	insert
SERV	reserve
SID	reside
SIGN#	resign
SIST	assist
SOCI	social
SOL	solid
solu(t) (SOLV9)	solution, insoluble
SOLV	dissolve
spair (SPER9)	despair
SPEC	species
SPECT	inspect
SPER	desperate
spic (SPECT9)	suspicion
SPIR	inspire
SPOND	respond
spons (SPOND9)	responsible
STAB	stable
stanc (STANT9)	instance
STANT	instant
STAT	statue
STINCT	distinct
stingu (STINCT9)	distinguish
STIT	substitute
stroy (STRUCT9)	destroy
STRUCT	instruct
SULT	insult
SUM#	assume
TAIL#	detail
TAIN	contain
TECT	detect
TEMP	temporary
ten (TAIN9)	tenant
TEND#	intend
tens (TEND#9)	intense
tent (TEND#9)	intention
tent (TAIN9)	content
TERR	territory
TEST#	protest
tin (TAIN9)	continent
TOR	torment
TRACT	tractor
TRIB	tribute
TURB	disturb
UN#	unite
VAC	vacant
VAL	value
VEL	develop
VEN	convenient
VENT	event
VERT	convert
vers (VERT9)	converse
VICT	victory
vic (VID9)	advice
VID	divide
vis (VID9)	division
vit (VID9)	invite
VIV	survive
VOC	vocal
vok (VOC9)	provoke
VOLV	revolve
vol(u)t (VOLV9)	revolution, revolt
WARD#	reward

Appendix 1

FREQUENCIES OF CORRESPONDENCES, AFFIXES, AND EPENTHETIC ELEMENTS

Table 1

Frequencies of Consonants

Grapheme Unit	Coding Number	Unmarked	Marked	Total
B		1110	%: 2	1112
	5	21		21
BB		34	+: 1	37
			%: 2	
C	1	74	+: 196	361
			+%: 14	
			@: 21	
			#: 56	
	2	1201	+: 21	1241
			%: 19	
	8	41		41
CC		9		9
CH		226	+: 13	239
	6	55		55
	7	15		15
CK		269	+: 7	276
D		1674	%: 2	1676
	9	16		16
DD		31	+: 4	36
			%: 1	
DG		25	+: 3	28
F		825	+: 2	863
			%: 36	
FF		19	+: 31	55
			%: 5	
G	1	21	+: 130	201
			+%: 3	
			@: 4	
			#: 43	
	2	632	%: 1	633
	5	15		15
	6	43		43
	9	5		5
GG		19	+: 2	21
GH		49		49
GUE		9		9
H		537		537
	5	30		30
J		174		174
K		500	+: 44	544
	5	33		33
L		1930	+: 38	1991
			%: 23	
	5	33		33
LL		104	+: 173	290
			%: 13	

Grapheme Unit	Coding Number	Unmarked	Marked	Total
M		1474	%: 34	1508
MM		22	%: 1	23
N		2181	%: 8	2189
	4	153		153
	5	7		7
	6	11		11
NG		93		93
NN		35	+: 7	53
			%: 11	
P		1656	%: 45	1701
PH		86		86
PP		28	%: 1	29
QU		85		85
	6	9		9
QUE		6		6
R		3318	%: 20	3338
RH		6		6
RR		101	+: 4	113
			%: 8	
S		1825	+: 16	1927
			%: 17	
			#: 69	
	2	64		64
	3	192	@: 22	214
	4	29		29
	6	73	@: 3	76
	7	13		13
	8	6		6
	9	37		37
SC		21		21
SH		267		267
SS		30	+: 108	155
			%: 10	
			@: 7	
	8	24		24
T		2855	%: 15	2870
	5	19		19
	8	247		247
	9	70		70
TCH		55		55
TH	1	204		204
	2	47		47
	3	32		32
TT		89	+: 9	101
			%: 3	

Table 1--continued

Grapheme Unit	Coding Number	Unmarked	Marked	Total
V		609	+: 1 @: 24 #: 47	681
W		456		456
	5	37		37
WH		64		64
X		85		85
	8	3		3
Y	1	53		53
Z		65	+: 3 @: 13 #: 1	82
ZZ		12	+: 7	19

Exceptional Correspondences

Grapheme Unit	Coding Number	Frequency	Total
B	45	8	8
BB	--	--	--
C	40	6	10
	45	4	
CC	40	3	3
CH	45	1	1
CK	--	--	--
D	40	1	4
	45	3	
DD	--	--	--
DG	40	1	1
F	40	4	4
FF	--	--	--
G	40	1	4
	41	2	
	45	1	
GG	--	--	--
GH	41	6	14
	42	8	
GUE	--	--	--
H	--	--	--
J	40	2	2
K	40	2	2
L	40	1	9
	45	8	
LL	40	2	2
M	40	2	2
MM	--	--	--
N	45	1	1
NG	41	6	6
NN	--	--	--
P	45	7	7
PH	40	1	1
PP	--	--	--
QU	--	--	--
QUE	--	--	--
R	40	4	5
	45	1	
RH	--	--	--
RR	--	--	--
S	40	13	45
	42	21	
	45	11	
SC	40	4	4
SH	--	--	--
SS	40	3	10
	41	7	
T	40	7	12
	45	5	
TCH	--	--	--
TH	41	17	40
	42	13	
	43	5	
	45	5	
TT	--	--	--
V	40	1	1
W	45	4	4
WH	42	12	12
X	40	7	7
Y	--	--	--
Z	40	4	4
ZZ	40	1	1

Table 2

Frequencies of Secondary Vowels

Grapheme Unit	Coding Number	Unmarked	#	Other Symbol	Total
AI		153	48		201
AU		68	5		73
AW		44	1	@: 7 #@:1	53
AY		107	1	@: 3	111
EA		243	41		284
	6	98	12		110
	7	19			19
EAU		3			3
EE		224	18		242
EI	1	20	3		23
	2	16	1		17
EU		4		+: 1 #+: 3	8
EW		57		+: 7 @+: 2	66
EY		15	1		16
	5	3			37
IE	1	7	1		8
	2	38	5		43
IGH		78			78
OA		83	24	@: 2	109
OE		13			13
OI		58			58
OO	1	151	5	@: 15	171
	2	90			90
OU		196	8	@: 3	207
	6	29	14		43
	7	19	13	+: 1 @: 6 @#: 1	40
	8	10			10
	9	14	18		32
OW	1	107		@: 7	114
	2	89		@: 8	97
OY		38		@: 3	41
UE		20		+: 7	27
UI		20			20
YE		9			9

Exceptional Correspondences

Grapheme Unit	Coding Number	Frequency	Total
AI	40	7	7
AU	40	5	10
	40	5	
AW	--	--	--
AY	40	11	11
EA	40	3	30
	41	16	
	42	11	
EAU	40	3	3
EE	40	2	2
EI	40	7	7
EU	40	1	1
EW	40	3	3
EY	40	4	13
	41	9	
IE	40	12	12
IGH	--	--	--
OA	40	3	8
	41	5	
OE	40	8	8
OI	40	3	3
OO	41	12	17
	42	5	
OU	40	8	8
OW	40	2	2
OY	40	1	1
UE	40	8	8
UI	40	[illegible]	10
	42	9	
YE	--	--	--
SV	40	18	18

Table 3

Frequencies of Primary Vowels

Grapheme Unit	11	11#	Total 11	12	13	13#	Total 13	14	15	16	16#	Total 16	17	17#	Total 17	21
A	346	56	402	21	102	53	155	10	694	428	--	428	330	32	362	229
E	29	15	44	--	70	22	92	4	573	282	37	319	196	83	279	129
I	327	48	375	9	60	14	74	12	619	301	1	302	258	9	267	66
O	199	56	255	4	175	41	216	4	247	144	--	144	205	32	237	233
U	95*	26*	121	2*	95*	14*	109	3	393	233	--	233	59*	18*	77	111
Y	8	--	8	4	5	2	7	8	10	8	--	8	15	--	15	--
Total	1004	201	1205	40	507	146	653	41	2536	1396	38	1434	1063	174	1237	838

	25	25#	Total 25	26	27	27#	Total 27	28	29	29#	Total 29	32	33	33#	33+	Total 33	39
A	129	--	129	1	31	1	32	6	108	--	108	5	83	--	16	99	11
E	16	1	17	48	10	6	16	6	91	15	106	26	62	12	8	82	1
I	22	--	22	41	23	3	26	--	65	2	67	5	111	1	4	116	1
O	99	--	99	8	29	--	29	6	60	--	60	2	39	--	8	47	8
U	--	--	--	1*	--	--	--	1	--	--	--	--	4	--	11	15	--
Y	72	--	72	2	2	--	2	--	4	1	5	--	--	1	2	3	5
Total	338	1	339	113	95	10	105	19	328	18	346	38	299	14	49	362	26

*Some correspondences for U are further coded to reflect the pronunciations /u/ and /ə/ (no code) or the pronunciations /yu/ and /yə/ ("+"): U11 (44), U11# (13), U11+ (51), U11#+ (13); U12+ (2); U13 (53), U13# (4), U13+ (42), U13#+ (10); U17 (37), U17# (14), U17+ (22), U17#+ (4); U26 (2), U26+ (11).

Table 3--continued

Grapheme Unit	Coding Number	Frequency	
A	18	31	
	20	33	
	22	31	
	23	83	
	24	47	
	34	37	
	35	14	
E	18	793	1386
	19	167	
	20	339	
	30	87	
	35	14	
I	18	45	
	35	9	
	36	34	
	37	12	
O	22	30	
	23	55	
	24	105	
	31	113	
	34	20	
U	20	17	
	24	14	
	31	31	
Y	36	2	

Exceptional Correspondences

Grapheme Unit	41	45	46	47	49
A	5	--	3	8	14
E	1	15	--	2	7
I	9	--	4	5	8
O	--	--	15	3	2
U	--	5	4	4	1
Y	--	--	--	--	1
Total	15	20	26	22	33

Grapheme Unit	Coding Number	Frequency
A	40	33
	48	5
E	40	8
	42	9
	43	13
	44	3
I	40	19
	42	10
O	40	27
	42	22
	43	14
	44	4
	48	13
U	40	11
Y	40	2

Total Frequency of Exceptional Correspondences	
A	68
E	58
I	55
O	100
U	25
Y	3
Total	309

Table 4

Frequencies of Prefixes

Code	Frequency	Total	Code	Frequency	Total	Code	Frequency	Total
A	30		EX	16		PRO	6	
A+	56	86	EX9	3		PRO+	1	
			EX2	57		PRO1	28	
AB	6		EX2+	1		PRO2	36	
AB2	9		EX92	8		PRO2+	1	72
AB2+	2	17	EX3	15	100			
						RE	5	
AD	9		FOR	1		RE+	24	
AD+	1		FOR+	11	12	RE1	35	
AD9	16					RE1+	7	
AD2	22		IN	28		RE2	104	
AD2+	3		IN+	5		RE2+	25	200
AD92	85		IN9	7				
AD92+	30		IN9+	1		SE	2	
AD3	4		IN2	57		SE1	7	
AD3+	2	172	IN2+	15		SE2	10	19
			IN92	17				
AL#+	6	6	IN92+	4	134	SUB	11	
						SUB+	3	
ANTI+	3		IN#	3		SUB9	26	
ANTI9	1		IN#+	9		SUB9+	1	41
ANTI9+	2	6	IN#9	2				
			IN#2	1		SUPER	4	
BE	23		IN#2+	9		SUPER+	6	10
BE+	15	38	IN#92	1				
			IN#92+	8	35	SUR	8	
CON	47					SUR+	5	13
CON+	6		INTER	8				
CON9	29		INTER+	4		SYN	3	
CON9+	5		INTER9	1		SYN9	7	10
CON91	7		INTER1	3				
CON2	80		INTER91	2		TELE	2	
CON2+	11		INTER2	3	21	TELE+	6	8
CON92	52							
CON92+	7		MIS	3		TRANS	11	
CON3	4	249	MIS+	6	9	TRANS+	3	
						TRANS9	1	15
DE	3		MONO+	2				
DE+	3		MONO9	1		UN+	76	76
DE1	24		MONO1	1				
DE2	81		MONO1+	1	5	PR	22	
DE2+	11	122				PR+	11	33
			MULTI	5	5			
DIA	5							
DIA1	2	7	OB	3				
			OB+	1				
DIS	2		OB9	11				
DIS+	17		OB2	12				
DIS9	12		OB92	11	38			
DIS2	20							
DIS2+	17		PARA	7				
DIS92	22		PARA+	3				
DIS92+	1		PARA1	1	11			
DIS3	2							
DIS3+	1	94	PER	7				
			PER+	1				
E	14		PER2	14	22			
E+	1							
E2	30		PRE	5				
E2+	10	55	PRE+	3				
			PRE1	13				
EN	8		PRE1+	1				
EN9	5		PRE2	21				
EN2	17		PRE2+	1	44			
EN2+	25							
EN92	11							
EN92+	2	68						

Table 5

Frequencies of Suffixes

Code	Frequency	Total	Code	Frequency	Total	Code	Frequency	Total
A#+	11		EE+	5	5	IC	77	
A#9+	1	12				IC+	41	
			EL	67		IC9	10	128
ABLE	6		EL+	3				
ABLE+	30		EL9	3		ICE	9	
ABLE9	2		EL9+	1	74	ICE+	6	
ABLE9+	3	41				ICE9	7	22
			EM	9	9			
ACLE	4					ID	21	
ACLE9	1	5	EN#	54		ID+	1	
			EN#+	71		ID9	1	23
ADE	8		EN#9	4				
ADE+	3	11	EN#9+	25	154	IE	22	
						IE+	15	
AGE	29		ENCE	21		IE9	1	38
AGE+	15		ENCE+	13				
AGE9	8		ENCE9	4	38	IFY	7	
AGE9+	1	53				IFY+	4	
			ENT	47		IFY9	9	
AIN	10		ENT+	20		IFY9+	2	22
AIN+	7		ENT9	11				
AIN9	2	19	ENT9+	5	83	IL	22	
						IL9	1	23
AL	96		ER	501				
AL+	96		ER+	403		ILE	12	
AL9	8		ER9	38		ILE9	6	18
AL9+	4	204	ER9+	3	945			
						IN@	58	
AN	33		ERN	10		IN@9	4	62
AN+	28		ERN+	13	23			
AN9	3					INE	5	
AN9+	1	65	ERY	11		INE9	1	
			ERY+	15		INE2	19	
ANCE	7		ERY9	4		INE2+	7	
ANCE+	24		ERY9+	1	31	INE3	14	
ANCE9	4	35				INE3+	5	51
			ESS	3				
ANT	31		ESS+	8	11	ING	14	
ANT+	13					ING+	353	367
ANT9	2	46	EST	14				
			EST9	2	16	IO	8	
AR	31					IO+	2	
AR+	16		ET	69		IO9	1	11
AR9+	1	48	ET+	6				
			ET9	2		ION	65	
ARD	18		ET2	4		ION+	218	
ARD+	2		ET2+	1	82	ION9	5	
ARD9+	1	21				ION1	10	
			ETTZ	3		ION1+	2	
ARY+	2		ETTE+	5	8	ION2+	2	302
ARY9	3							
ARY2	4		FUL	3		IS	25	
ARY2+	3		FUL+	54	57	IS+	4	
ARY3	17					IS9	2	31
ARY3+	11	40	HOOD+	5	5			
						ISH	37	
ATE	131		IA	6		ISH+	16	
ATE+	32		IA+	1		ISH9	1	54
ATE2	36		IA9	6				
ATE2+	16		IA1	10		ISM	2	
ATE92	3	218	IA1+	2		ISM+	6	8
			IA2	12				
ED+	38		IA2+	2	39	IST	7	
ED9+	20					IST+	8	
ED2+	127		IBLE	8		IST9	1	
ED3+	89		IBLE+	4		IST9+	1	17
ED4	4		IBLE9	3				
ED4+	9	287	IBLE9+	1	16	IT@	26	
						IT@+	4	30

Table 5--continued

Code	Frequency	Total	Code	Frequency	Total
ITE	8		OUS	39	
ITE+	1		OUS+	37	
ITE9	2		OUS9	2	78
ITE9+	1				
ITE2	1		RY	7	
ITE2+	6	19	RY+	5	12
ITY	13		S11	11	
ITY+	35		S11+	49	
ITY9	2		S12	21	
ITY9+	3	53	S12+	85	
			S13	2	
IVE	11		S13+	4	
IVE+	34		S14+	19	191
IVE9	1				
IVE9+	1	47	SHIP	1	
			SHIP+	12	13
IZE	6				
IZE+	14		SOME+	5	5
IZE9+	2	22			
			STER	2	
LE	224		STER+	3	5
LE+	20				
LE9	52		T+	7	7
LE9+	4	300			
			TEEN+	14	14
LESS	1				
LESS+	24	25	TH+	36	36
LET	1		TY	8	
LET+	6	7	TY+	14	
			TY9	5	27
LY	5				
LY+	254		UDE	6	
LY9+	1	260	UDE+	1	7
MENT	22		ULE	4	
MENT+	32		ULE9	1	5
MENT9	8				
MENT9+	1	83	UM	25	
			UM+	2	27
NESS	1				
NESS+	27	28	UNE	9	9
OM	18		URE	39	
OM9	1	19	URE+	16	
			URE9	1	
ON	89		URE2	2	
ON+	1		URE2+	1	59
ON9	2				
ON2	14	106	US	26	
			US+	1	27
OP	6	6			
			UTE	10	
OR	80		UTE9	1	
OR+	55		UTE2	10	21
OR9	9				
OR9+	2	146	WARD	1	
			WARD+	20	
ORY	4		WARD9+	2	23
ORY+	3				
ORY9	7		Y	256	
ORY9+	1		Y+	166	
ORY2	6		Y9	4[illegible]	
ORY2+	3	24	Y9+	10	472
OT	11				
OT9	1	12	END	15	
			END+	40	55

Table 6

Frequencies of Epenthetic Elements

Code	Frequency
I-	62
I1-	55
I2-	17
I3-	7
U-	24
U2-	9
AT-	11
AT1-	51
IT-	31
IT1-	19
IC-	11
IC1-	5
UL-	11
O,	21
O1,	10
I,	14
I1,	3